Department of Computer Science
Series of Publications A
Report A-2006-4

Algorithms for ^{13}C metabolic flux analysis

Ari Rantanen

To be presented, with the permission of the Faculty of Science of the University of Helsinki, for public criticism in Auditorium CK112, Exactum, on November 22th, 2006, at noon.

University of Helsinki
Finland

Contact information

Postal address:
Department of Computer Science
P.O. Box 68 (Gustaf Hällströmin katu 2b)
FI-00014 University of Helsinki
Finland

Email address: postmaster@cs.Helsinki.FI (Internet)

URL: http://www.cs.Helsinki.FI/

Telephone: +358 9 1911

Telefax: +358 9 191 51120

ISSN 1238-8645
ISBN 952-10-3510-2 (paperback)
ISBN 952-10-3511-0 (PDF)
Computing Reviews (1998) Classification: G.2.1, G.2.2, I.6.5, J.3
Helsinki 2006
Helsinki University Printing House

Algorithms for ^{13}C metabolic flux analysis

Ari Rantanen

Department of Computer Science
P.O. Box 68 (Gustaf Hällströmin katu 2b)
FIN-00014 University of Helsinki, Finland
ari.rantanen@cs.helsinki.fi

PhD Thesis, Series of Publications A, Report A-2006-4
Helsinki, November 2006, 92 + 73 pages
ISSN 1238-8645
ISBN 952-10-3510-2 (paperback)
ISBN 952-10-3511-0 (PDF)

Abstract

The metabolism of an organism consists of a network of biochemical reactions that transform small molecules, or metabolites, into others in order to produce energy and building blocks for essential macromolecules. The goal of metabolic flux analysis is to uncover the rates, or the fluxes, of those biochemical reactions. In a steady state, the sum of the fluxes that produce an internal metabolite is equal to the sum of the fluxes that consume the same molecule. Thus the steady state imposes linear balance constraints to the fluxes. In general, the balance constraints imposed by the steady state are not sufficient to uncover all the fluxes of a metabolic network. The fluxes through cycles and alternative pathways between the same source and target metabolites remain unknown.

More information about the fluxes can be obtained from isotopic labelling experiments, where a cell population is fed with labelled nutrients, such as glucose that contains ^{13}C atoms. Labels are then transferred by biochemical reactions to other metabolites. The relative abundances of different labelling patterns in internal metabolites depend on the fluxes of pathways producing them. Thus, the relative abundances of different labelling patterns contain information about the fluxes that cannot be uncovered from the balance constraints derived from the steady state. The field of research that estimates the fluxes utilizing the measured constraints to the relative abundances of different labelling patterns induced by ^{13}C labelled nutrients is called ^{13}C metabolic flux analysis.

There exist two approaches of ^{13}C metabolic flux analysis. In the optimization approach, a non-linear optimization task, where candidate fluxes are iteratively generated until they fit to the measured abundances of different labelling patterns, is constructed. In the direct approach, linear balance constraints given by the steady state are augmented with linear constraints derived from the abundances of different labelling patterns of metabolites. Thus, mathematically involved non-linear optimization methods that can get stuck to the local optima can be avoided. On the other hand, the direct approach may require more measurement data than the optimization approach to obtain the same flux information. Furthermore, the optimization framework can easily be applied regardless of the measurement technology and with all network topologies.

In this thesis we present a formal computational framework for direct ^{13}C metabolic flux analysis. The aim of our study is to construct as many linear constraints to the fluxes from the ^{13}C labelling measurements using only computational methods that avoid non-linear techniques and are independent from the type of measurement data, the labelling of external nutrients and the topology of the metabolic network. The presented framework is the first representative of the direct approach for ^{13}C metabolic flux analysis that is free from restricting assumptions made about these parameters. In our framework, measurement data is first propagated from the measured metabolites to other metabolites. The propagation is facilitated by the flow analysis of metabolite fragments in the network. Then new linear constraints to the fluxes are derived from the propagated data by applying the techniques of linear algebra. Based on the results of the fragment flow analysis, we also present an experiment planning method that selects sets of metabolites whose relative abundances of different labelling patterns are most useful for ^{13}C metabolic flux analysis. Furthermore, we give computational tools to process raw ^{13}C labelling data produced by tandem mass spectrometry to a form suitable for ^{13}C metabolic flux analysis.

Computing Reviews (1998) Categories and Subject Descriptors:
G.2.1 Combinatorics: Combinatorial algorithms
G.2.2 Graph theory: Graph algorithms
I.6.5 Model development
J.3 Life and medical sciences: Biology and genetics

General Terms:
Algorithms, Bioinformatics, Computational biology, Systems biology

Additional Key Words and Phrases:
Flow analysis, Isotopomer analysis, Mass spectrometry, Metabolic flux analysis, Metabolic modelling

Acknowledgements

This thesis would have never seen the light of day without the comprehensive guidance of my supervisors Esko Ukkonen and Juho Rousu. I am grateful for their help in every step of the process, from my first day as graduate student to the proof reading of the manuscript of the thesis. I thank Jaakko Hollmén and Sampsa Hautaniemi for their profound review of the thesis and their very helpful comments.

The contributions of this thesis are results of the team work. My deepest thanks go to Esa Pitkänen, who has — in addition to significantly contributing to the ideas of the thesis — tirelessly implemented and supervised the implementation of the multitude of software components that were necessary to test the ideas in practice. A collaboration with Taneli Mielikäinen has shown me creative science at its best. Paula Jouhten has given me invaluable insight to common practices in metabolic modelling as well as to actual biological processes behind the formal models. The solid performance of Markus Heinonen and Arto Åkerlund in the development and the implementation of the components of the thesis has been most important. Numerous discussions with Hannu Maaheimo, Juha Kokkonen and Raimo Ketola have been essential in maturing the ideas of the thesis. Katja Saarela's work was indispensable to get things going. I also thank Esa, Hannu, Paula and Marina Kurtén for their comments on the manuscript of the thesis.

I thank Matti Kääriäinen and Teemu Kivioja for their expert consultation. Janne, Kimmo, Pasi, Pekko and Veli deserve thanks for their intellectual support.

The research whose results are described in the thesis has been carried out at the Department of Computer Science of the University of Helsinki. I thank the department, especially its computing facilities staff, for a wonderful working environment. The financial support from the Academy of Finland, TEKES, FDK research unit and ComBi graduate school is greatly appreciated.

I owe my parents a great debt for their support. This is what can

happen if you buy a child a computer instead of a game console. My greatest gratitude goes to my love Katri, who has held the fort at home while I have worked long hours at the office. Elias and Ruu, I am sorry I have been away so much.

Contents

Original Publications

This thesis is based on the following publications, which are referred to in the text by their Roman numerals, and on unpublished results presented in the introductory Part I of the text.

I Juho Rousu, Ari Rantanen, Hannu Maaheimo, Esa Pitkänen, Katja Saarela, Esko Ukkonen:
A method for estimating metabolic fluxes from incomplete isotopomer information.
Proceedings of International Workshop on Computational Methods in Systems Biology, Rovereto Italy, February 2003. Lecture Notes in Computer Science 2602 (2003), pp. 88–103.

II Ari Rantanen, Hannu Maaheimo,Esa Pitkänen, Juho Rousu, Esko Ukkonen:
Equivalence of metabolite fragments and flow analysis of isotopomer distributions for flux estimation.
Transactions on Computational Systems Biology, Vol. 1, Lecture Notes in Bioinformatics 4220 (2006), pp. 198–220.

III Ari Rantanen, Taneli Mielikäinen, Juho Rousu, Hannu Maaheimo, Esko Ukkonen:
Planning optimal measurements of isotopomer distributions for estimation of metabolic fluxes.
Bioinformatics, Vol. 22, Number 10 (2006), pp. 1198–1206.

IV Ari Rantanen, Juho Rousu, Juha T. Kokkonen, Virpi Tarkiainen, Raimo A. Ketola:
Computing Positional Isotopomer Distributions from Tandem Mass Spectrometric Data.
Metabolic Engineering, Vol. 4 (2002), pp. 285–294.

V Juho Rousu, Ari Rantanen, Raimo A. Ketola, Juha T. Kokkonen:
Isotopomer distribution computation from tandem mass spectrometric data with overlapping fragment spectra.
Spectroscopy, Vol. 19 (2005), pp. 53–67.

Part I

Mathematical notations for Part I

$\mathcal{M}_i$	Metabolite with index i
$M_i = \{c_1, \ldots, c_k\}$	Set of carbon locations of metabolite $\mathcal{M}_i$
$\rho_j = (\alpha_j, \lambda_j)$	Biochemical reaction with index j, stoichiometric coefficients α_j and carbon mapping λ_j
$G = (\mathcal{C}, \mathcal{R})$	Metabolic network, where $\mathcal{C} = \{M_1, \ldots, M_m\}$ and $\mathcal{R} = \{\rho_1, \ldots, \rho_n\}$
$M\|F$	Fragment F of M, that is, a subset of carbons in M
$M(b)$	Set of molecules that belong to b-isotopomer of M, $b = (b_1, \ldots, b_k) \in \{0,1\}^k$, where $b_i = 0$ denotes a ^{12}C and $b_i = 1$ denotes a ^{13}C in location c_i
$M(+p)$	Set of molecules that belong to mass isotopomer $+p$ of M, that is, molecules of M that have p ^{13}C labels
$P_M(b)$	Relative abundance of the isotopomer b in M
$D(M)$	Isotopomer distribution of M
$\mathcal{I}_M$	Isotopomer space of M
$d_{i,h}$	Relative abundance of linear combination h of the isotopomers of M_i
$\iota_j^{k,l}$	Isotopomer mapping from the isotopomer space of substrate fragment $M\|F_k$ of ρ_j to the isotopomer space of product fragment $M'\|F_l$ of ρ_j
$IMM_j^{k,l}$	Isotopomer mapping matrix from substrate fragment $M\|F_k$ of ρ_j to product fragment $M'\|F_l$ of ρ_j
β_i	Measured external inflow or outflow of M_i
M_{ij}	Subpool of M_i produced or consumed by ρ_j
M_{i0}	Subpool of M_i that is related to the external inflow or outflow
v_j	Flux of reaction ρ_j
$\mathbf{v} = [v_1, \ldots, v_n]$	Flux distribution
$\mathcal{F}(G)$	Fragment flow graph of metabolic network G
T	Dominator tree of $\mathcal{F}(G)$
$idom(F)$	Immediate dominator of fragment F
$\otimes$	Component-wise Kronecker product

Chapter 1

Introduction

This thesis presents novel algorithms for ^{13}C metabolic flux analysis. The thesis belongs to the field of computational biology where "data-analytical methods, mathematical modelling and computational simulation techniques are developed and applied to study biological, behavioral, and social systems" [HDH+00]. The thesis is also a part of systems biology, "the science of discovering, modeling, understanding and ultimately engineering at the molecular level the dynamic relationships between the biological molecules that define living organisms" [Hoo].

The thesis consists of two parts, Part I and Part II. The main contributions are presented in the five publications constituting Part II. The aim of the introductory Part I is to associate these contributions to the full process of ^{13}C metabolic flux analysis and compare the contributions to existing methods. This first chapter of Part I briefly discusses the practical importance of metabolic flux analysis and then lists author's contributions to the subject. Chapter 2 formally defines the basic concepts used throughout the thesis, introduce the concept of stoichiometric modelling of metabolic networks and the measurement technologies relevant to the thesis. In Chapter 3 common assumptions behind ^{13}C metabolic flux analysis are listed and existing computational methods are reviewed. In Chapter 4 a process for ^{13}C metabolic flux analysis is proposed. Chapter 5 discusses the preprocessing of measurement data for ^{13}C metabolic flux analysis. Chapter 6 concludes Part I and sketches some directions for future work. The sections of Part I denoted with "*" contain technical discussion that can be skipped without great loss of continuity.

1.1 Metabolic fluxes and the program of life

One of the most intriguing open questions in modern natural science is to understand operational principles of living organisms. We know that most functions sustaining life are executed by proteins that are molecules consisting of chains of amino acids [AJL+02]. We also know that the instructions for building proteins are coded to double-stranded DNA molecules with a four-letter alphabet. The wealth of genome mapping projects continue to provide us with these codes for different organisms [BKML+04], including ourselves [Lea04]. We understand the processes of RNA and protein syntheses that transform the genetic information stored to DNA into proteins. For many proteins, the genes coding them are known [BKML+04] and for many – but not for all – of them also some function is annotated [Bea05]. But still the operational principles, or "the program" of life, escapes our comprehensive understanding. Knowing the DNA of an organism does not decipher this program, it only gives us a coded list of parts used to construct an immensely complex system – the system-wide mechanisms that regulate the production of proteins and thus control the execution of the program of life are still incompletely understood. Comparing the situation to computer programming, we only have fragmentary and inaccurate knowledge about the basic primitives (proteins) of a programming language used to implement a very complex system but the control flow of the program is largely unknown to us.

The difficulty of understanding the program of life stems from the fact that neither the source code of the program nor the syntax of the programming language are directly readable. To study an organism as a complete system [Kit02] we can only perturb it and monitor its responses, that is, read the outputs of the program when different inputs are given to it and some parts of the code are (randomly or systematically) altered [IGH01]. The difficulties of this kind of an approach can be understood by thinking of an analogous method for understanding the operational principles of a radio, presented by Lazebnik [Laz04]: first a huge amount of working radios are built. Then the radios are shot with a gun and the components that were hit in malfunctioning samples are identified as essential parts that should get all the attention in further studies.

The successful application of such a "knowledge through perturbation" method requires, among other things, a good modelling language to describe the hypotheses about the behaviour of the system [Laz04]. It should also be helpful to be able to monitor the responses of an organism, or phenotype, from all relevant points of view [GWV03]. Nowadays the phenotype of the perturbed subject of experiment can be investigated in different "omics"

levels. For example, in transcriptome profiling the abundances of mRNA transcripts produced by RNA synthesis can be simultaneously measured for thousands of genes [ESBB98, LW00]. Similarly, in proteome profiling at least qualitative information on hundreds, even thousands of proteins can be obtained [WWJ01]. Protein–protein, protein–DNA and protein–RNA interactions, or interactome, of an organism can also be studied with high throughput methods [Fea99, Hea02].

1.1.1 Metabolic fluxes are an important phenotype

Recently, the study of the *metabolism* has given us a chance to gain information on the phenotype of an organism from a novel point of view [Fie02, FTKL04, FGS05]. The metabolism of a living cell consists of biochemical reactions transforming small molecules, metabolites to others by cleaving and combining them. The reactions of a metabolism are interconnected through common metabolites and thus form metabolic networks where the products of one reaction act as substrates for another reaction. Figure (1.1) depicts an example of a metabolic network.

Through its metabolism, an organism performs two fundamental tasks [BTS02, AJL+02]:

1. Generation of energy by breaking down nutrient molecules,

2. Synthesization of building blocks of macromolecules, such as amino acids, and eventually macromolecules themselves.

The metabolic reactions are significantly speeded up, or *catalyzed* by enzymes, proteins that bind to substrates and lower the activation energy of the reactions [BTS02]. The velocity, or the *flux*, of a metabolic reaction depends on the properties of enzymes catalyzing the reaction, and concentrations of substrates, products and other metabolites affecting the activity of catalyzing enzymes. The concentrations of enzymes depend on the rate of RNA and protein synthesis and degradation while the concentrations of metabolites depend on the fluxes of reactions producing and consuming them. By producing different amounts of enzymes at different times an organism can regulate its fluxes and adapt to different conditions by building and breaking molecules most appropriate for the situation. Thus metabolite levels and metabolic fluxes, or *metabolome* and *fluxome*, can be seen as "the ultimate" phenotype of an organism to genetic or environmental changes [Fie02, Nie03]. Specifically, "metabolic fluxes constitute a fundamental determinant of cell physiology because they provide a measure of the degree of engagement of various pathways in overall cellular function

and metabolic processes" [SAN98]. While the study of the steady state metabolic fluxes alone is not enough to decode the program of life, when combined with other types of information, they can give important insight to the operational principles of an organism and its capabilities to adapt to different conditions and help us to understand the function of genes involving metabolic regulation [Nie03, WvWvGH05].

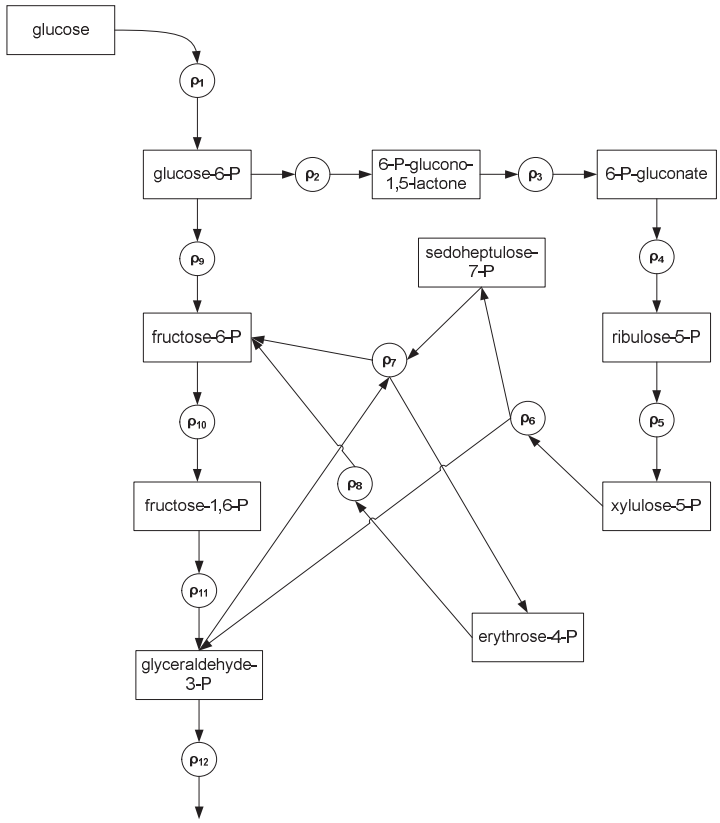


Figure 1.1: A part of the metabolic network of *Saccharomyces cerevisiae* [BKS05]. Rectangles represent metabolites and circles reactions.

Currently, the metabolic fluxes are mostly analyzed in the field of metabolic engineering, where microbial organisms are genetically modified to improve the product formation or cellular properties [SAN98]. System-wide flux information revealing the degree of the activity of metabolic pathways can be utilized e.g. in the comparison of

1. the phenotypes of an organism in different environmental conditions

[FW05, GMdSCN01, SMY$^+$04],

2. different genetic strains of an organism [BKS05, EDP$^+$02, GCNO05],
3. related species [BLS05], and
4. *in vivo* and *in vitro* behaviour of an enzyme [SAN98].

In addition to microbes, flux analysis of plants [RSH06] can applied with analogous goals. In the study of mammalian cells, the information about the metabolic fluxes can help in better understanding of diseases [TK96, Hel03] and in more efficient drug design [BSCL04, Tur06].

1.2 ^{13}C metabolic flux analysis

In a steady state, the sum of the fluxes that produce an internal metabolite is equal to the sum of the fluxes that consume the same molecule (see Section 2.2). Thus, the steady state imposes linear balance constraints to the fluxes. However, the balance constraints imposed by the steady state are not sufficient to uncover all the fluxes of a metabolic network. The fluxes through cycles, backward fluxes and the fluxes through alternative pathways between source and target metabolites remain unknown.

More constraints to the fluxes can be obtained from isotopic labelling experiments. In the isotopic labelling experiments a cell population is cultivated with labelled nutrients, such as glucose that contains ^{13}C atoms (Section 2.3). Biochemical reactions then transfer the nutrient labels to other metabolites in the network.

Different metabolic pathways manipulate the carbon chains of metabolites in their characteristic ways and thus induce different kinds of labelling patterns to their metabolites. The relative abundances of different labelling patterns in metabolites depend on the fluxes of pathways producing them. Thus, the relative abundances of different labelling patterns contain information about the fluxes that is not present in the balance constraints derived from the steady state. The abundances of different labelling patterns — or constraints to them — can be measured either by mass spectrometry (MS) or by nucleic magnetic resonance spectroscopy (NMR) (Section 2.4). The field of research that estimates the fluxes utilizing the measured constraints to the relative abundances of different labelling patterns induced by ^{13}C labelled nutrients is called ^{13}C metabolic flux analysis. At a high level, the process of ^{13}C metabolic flux analysis consists of the following steps: First, the model of a metabolic network is constructed. Then, a cell population is cultivated with labelled nutrients and the abundances of different

labelling patterns in metabolites are measured. Next, the raw measurement data is preprocessed to the form that is suitable for ^{13}C metabolic flux analysis (Chapter 5). Finally, utilizing both the model of the metabolic network and the preprocessed measurement data, metabolic fluxes are estimated. A more detailed description of the process of ^{13}C metabolic flux analysis proposed in this thesis is given in Chapter 4.

There exist two general approaches for ^{13}C metabolic flux analysis (Section 3.3) that differ in computational methods employed in the flux estimation step. In the *optimization approach*, fluxes are estimated by constructing and solving a non-linear optimization task, where candidate fluxes are iteratively generated until they fit to the measured abundances of different labelling patterns. In the *direct approach*, linear balance constraints given by the steady state are augmented with linear constraints derived from the abundances of different labelling patterns of metabolites. Thus, mathematically involved non-linear optimization methods that can get stuck to the local optima can be avoided. On the other hand, the direct approach may require more measurement data than the optimization approach to obtain the same flux information. Also, the optimization framework can be easily applied regardless of the quality of the ^{13}C labelling measurements and with all network topologies.

1.3 Contributions

This thesis presents a formal computational framework for direct ^{13}C metabolic flux analysis. The aim of our study is to construct a largest possible number of linear constraints to the fluxes from the ^{13}C labelling measurements using only computational methods that avoid non-linear techniques and are independent from the quality of measurement data, the labelling of external nutrients and the topology of the metabolic network.

The main contributions of this thesis are given in five publications constituting Part II. In Publication I we introduce a general framework for ^{13}C metabolic flux analysis where incomplete isotopomer measurements are interpreted as linear constraints to the isotopomer distributions of metabolites. These linear constraints are propagated from the measured metabolites to unmeasured ones. From the constraints to the isotopomer distributions of metabolites linear constraints to the flux distribution are then inferred. Together with stoichiometric constraints, these flux constraints form a linear equation system that is then solved to obtain an estimate of the complete flux distribution. The framework of Publication I can be applied to all network topologies and all isotopomer distributions of

input substrates and can simultaneously take advantage of isotopomer information produced by mass spectrometry or by nucleic magnetic resonance spectroscopy.

Publication II gives an efficient algorithm to partition the fragments of metabolites in the network to equivalence classes that have equal isotopomer distributions in every steady state. This partition facilitates a more efficient method for propagating measured isotopomer information in the metabolic network than the propagation method of Publication I. Together, fragment equivalence classes and the framework of Publication I generalize and formalize existing METAFoR methods for ^{13}C metabolic flux analysis [Szy95, SGH$^+$99, MFC$^+$01] that assume uniform labelling of input substrates and compute only local ratios of fluxes producing the same metabolite. The framework of Publication I and the fragment equivalence classes also generalize the methods of ^{13}C constrained flux balancing where mass balances and flux ratios are combined to obtain the complete flux distribution, but that are bound to certain measurement techniques and input substrate labellings, such as uniform labelling of substrates and NMR data [SHB$^+$97] or MS data [FNS04]. Fragment equivalence classes also facilitate methods for structural identifiability analysis and for improving the noise tolerance of flux estimations, as described in Part I.

The measurement of isotopomer distributions of internal metabolites is a tedious and non-trivial task. Thus, it is worthwhile to concentrate the measurement efforts to metabolites that are most useful for ^{13}C metabolic flux analysis, that is, to subsets of metabolites whose isotopomer distributions give enough information to uncover the fluxes. With fragment equivalence classes and certain assumptions about the quality of the measurement data, the selection of most informative metabolites to measure can be formulated as a variant of the classical set cover problem. The experiment planning algorithms for selecting metabolites to measure are given in Publication III.

Publication IV and Publication V describe algorithms to preprocess raw data produced by tandem mass spectrometry (MS-MS) to a form suitable for ^{13}C metabolic flux analysis. Publication IV extends the method of Christensen and Nielsen [CN99] for computing constraints to the isotopomer distribution of a metabolite from data produced by GC-MS with full scanning fragmentation method (see Section 2.4): the method of Publication IV can also be applied when MS-MS with daughter ion scanning is used to fragment metabolite molecules. Compared to the full scanning technique, daughter ion scanning has a potential to produce complementary constraints to the isotopomer distribution of a metabolite. Thus the

contribution of Publication IV can help in ^{13}C metabolic flux analysis. Publication V extends the method of Publication IV to utilize also information in overlapping daughter ion spectra to compute even more constraints to the isotopomer distributions of metabolites from MS-MS data.

Introductory Part I contains the following new contributions that generalize some results given in Part II to the complete computational process for ^{13}C metabolic flux analysis described in Chapter 4 of Part I. In Section 4.4.1, upper bounds to flux information obtainable from isotopomer balance equations constraining the fluxes (see Section 2.3) are derived. Then, in Section 4.4, the upper bounds are utilized in structural identifiability analysis [IW03, vWHVG01], which studies, whether available measurements can in principle give enough information to fix the values of the fluxes in the network. Furthermore, in Section 4.7.4 we show how the upper bounds to the flux information can be used to improve the tolerance of the proposed flux analysis method to experimental errors. Another analysis technique of fragment equivalence classes to improve the propagation of measurement data is given in Section 4.3.2.

For completeness, an unpublished software for constructing metabolic network models for ^{13}C metabolic flux analysis and a computational method for identifying metabolite fragments produced by MS-MS [HRM$^+$06] are shortly described in Sections 4.2 and 5.1.

The results reported in the thesis were obtained, often in very close collaboration, by the author and the other members of the computational systems biology research group, lead by Juho Rousu and Esko Ukkonen. The ideas behind publications I and V were developed jointly by the author and Juho Rousu. The author implemented the methods of Publication I and co-designed and conducted the experiments reported in the publication. The author supervised the implementation and partly implemented the method and conducted the computational experiments described in Publication V. The main technical ideas behind Publications II, III and IV are due to the author. The author also implemented the methods of these publications and designed and conducted the computational experiments reported in the publications. The MILP program described in Section 4.3 of Publication III is co-designed by Taneli Mielikäinen and the author. The author participated in the writing of all the papers.

The new results reported in Part I are due to the author with the exception of the software for constructing metabolic network models (Section 4.2) which was developed jointly by the author, Esa Pitkänen, and Arto Åkerlund. In particular, the author designed and implemented the software for metabolic flux estimation described in Sections 4.3 – 4.5 and

4.7 of Part I as well as designed and conducted the flux analysis reported in Section 4.8. The (unpublished) isotopomer data for the analysis was provided by VTT. The model of the metabolism of *Saccharomyces cerevisiae* used in Section 4.8 was established by Paula Jouhten and Hannu Maaheimo.

Chapter 2

Preliminaries

In this chapter we formally define basic concepts used throughout the thesis. Then we introduce the stoichiometric modelling of metabolic networks, the use of ^{13}C labelling data to uncover information about the metabolic fluxes and the measurement technologies for obtaining ^{13}C labelling data.

2.1 Formal definitions

In ^{13}C metabolic flux analysis the carbon atoms of metabolites are of special interest. Thus we usually represent a k-carbon *metabolite* $\mathcal{M}$ as a set of carbon locations $M = \{c_1, \ldots, c_k\}$. For simplicity, also M is called metabolite, when only carbons are of interest. A *metabolic network* $G = (\mathcal{C}, \mathcal{R})$ is composed of a set $\mathcal{C} = \{M_1, \ldots, M_m\}$ of metabolites and a set $\mathcal{R} = \{\rho_1, \ldots, \rho_n\}$ of *reactions* that perform the interconversions of metabolites. Here reaction $\rho \in \mathcal{R}$ represents a sum total of cellular reactions of the same kind in the network and metabolite $M \in \mathcal{C}$ a *pool* of metabolite molecules that have the same molecular structure. *Fragments* of metabolites are subsets $F = \{f_1, \ldots, f_h\} \subseteq M$ of the metabolite. A fragment F of M is denoted as $M|F$. Metabolites that are taken up into the cell from the growth medium are called *external substrates* or *external nutrients*.

With *isotopomers* we mean molecules with similar element structure but different combinations of ^{13}C labels (see Figure 2.1). Isotopomers of $M = \{c_1, \ldots, c_k\}$ are represented by binary sequences $b = (b_1, \ldots, b_k) \in \{0,1\}^k$ where $b_i = 0$ denotes a ^{12}C and $b_i = 1$ denotes a ^{13}C in location c_i. Molecules that belong to the *b–isotopomer* of M are denoted by $M(b)$. Isotopomers of metabolite fragments $M|F$ are defined in an analogous manner: a molecule belongs to the $F(b)$*–isotopomer* of M, denoted $M|F(b_1, \ldots, b_h)$, if it has a ^{13}C atom in all locations f_j that have $b_j = 1$, and ^{12}C in other

locations of F. Isotopomers with equal numbers of labels belong to the same *mass isotopomer*. We denote *mass isotopers* of M by $M(+p)$, where $p \in \{0, \ldots, |M|\}$ denotes the number of labels in isotopomers belonging to $M(+p)$.

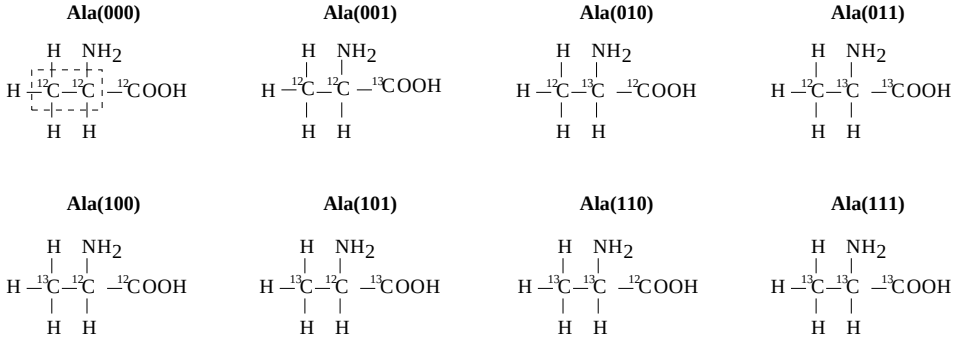


Figure 2.1: Eight possible isotopomers of alanine. The mass isotopomers are: $Ala(+0) = \{Ala(000)\}$; $Ala(+1) = \{Ala(001), Ala(010), Ala(001)\}$; $Ala(+2) = \{Ala(011), Ala(101), Ala(110)\}$; $Ala(+3) = \{Ala(111)\}$. In $Ala(000)$, carbons enclosed by a rectangle constitute a fragment.

The *isotopomer distribution* $D(M)$ of metabolite M gives the relative abundances $0 \leq P_M(b) \leq 1$ of each isotopomer $M(b)$ in the pool of M such that

$$\sum_{b \in \{0,1\}^{|M|}} P_M(b) = 1.$$

The isotopomer distribution $D(M|F)$ of fragment $M|F$ and the *mass isotopomer* distribution $D(M)^m$ of mass isotopomers $M(+p)$ are defined analogously: $D(M|F)$ of metabolite M gives the relative abundances $0 \leq P_{M|F}(b) \leq 1$ of each isotopomer $M|F(b)$ and $D(M)^m$ gives the relative abundances $0 \leq P_M(+p) \leq 1$ of each mass isotopomer $M(+p)$. By $d_{i,h}$ we denote the relative abundance of linear combination h of isotopomers of M_i (the concept is elaborated in Section 2.4).

Reactions are pairs $\rho_j = (\alpha_j, \lambda_j)$ where $\alpha_j = (\alpha_{1j}, \ldots, \alpha_{mj}) \in \mathbb{Z}^m$ is a vector of *stoichiometric coefficients*—denoting how many molecules of each kind are consumed and produced in a single reaction event—and λ_j is a carbon mapping describing the transition of carbon atoms in ρ_j (see Figure 2.2). If $\alpha_{ij} < 0$, a reaction event of ρ_j consumes $|\alpha_{ij}|$ molecules of M_{ij}, and if $\alpha_{ij} > 0$, it produces $|\alpha_{ij}|$ molecules of M_i. Metabolites M_i with $\alpha_{ij} < 0$ are called *substrates* and metabolites with $\alpha_{ij} > 0$ are called

products of ρ_j. If a metabolite is a product of at least two reactions, it is called a *junction*.

In the following, we assume that the reactions have *simple stoichiometries* $\alpha_{ij} \in \{-1, 0, 1\}$ and that the carbon mappings λ_j are bijections. Reactions producing or consuming many copies of the same metabolite molecules or symmetric metabolites can be modelled using simple stoichiometries by a simple transformation given in Section 5 of Publication II. Bidirectional reactions are modelled as a pair of reactions.

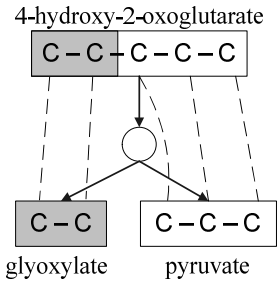


Figure 2.2: An example of a metabolic reaction. In 4-hydroxy-2-oxoglutarate glyoxylate-lyase reaction a 4-hydroxy-2-oxoglutarate ($C_5H_6O_6$) molecule is split into pyruvate ($C_3H_4O_3$) and glyoxylate ($C_2H_2O_3$) molecules. Carbon maps are shown with dashed lines (figure from Publication II).

A *pathway* in network G from metabolite fragments $\{F_1, \ldots, F_p\}$ to fragment F' is a sequence of reactions that define a (composite) mapping from the carbons of $\{F_1, \ldots, F_p\}$ to the carbons of F'.

It will be useful to distinguish between the *subpools* of a metabolite pool produced by different reactions. Therefore, we denote by M_{ij}, the subpool of the pool of M_i produced ($\alpha_{ij} > 0$) or consumed ($\alpha_{ij} < 0$) by reaction ρ_j.

By M_{i0} we denote the subpool of M_i that is related to the external inflow or external outflow of M_i. We call the sources of external inflows *external substrates*. Subpools of fragments are defined analogously.

In ^{13}C metabolic flux analysis, the quantities of interest are the rates or the fluxes $v_j \geq 0$ of the reactions ρ_j, giving the number of reaction events of ρ_j per time unit. We denote by $\mathbf{v}$ the vector $[v_1, \ldots, v_n]$ of fluxes. Slightly abusing terminology, $\mathbf{v}$ is often called a *flux distribution*.

	ρ_1	ρ_2	ρ_3	ρ_4	ρ_5	ρ_6	ρ_7	ρ_8	ρ_9	ρ_{10}	ρ_{11}	ρ_{12}
glucose	-1	0	0	0	0	0	0	0	0	0	0	0
glucose-6-P	1	-1	0	0	0	0	0	0	-1	0	0	0
6-P-G-1,5-L	0	1	-1	0	0	0	0	0	0	0	0	0
6-P-gluconate	0	0	1	-1	0	0	0	0	0	0	0	0
ribulose-5-P	0	0	0	1	-1	0	0	0	0	0	0	0
xylulose-5-P	0	0	0	0	1	-2	0	0	0	0	0	0
S-7-P	0	0	0	0	0	1	-1	0	0	0	0	0
erythrose-4-P	0	0	0	0	0	0	1	-1	0	0	0	0
fructose-6-P	0	0	0	0	0	0	1	1	1	-1	0	0
fructose-1,6-P	0	0	0	0	0	0	0	0	0	1	-1	0
G-3-P	0	0	0	0	0	1	-1	0	0	0	1	-1

Table 2.1: The stoichiometric matrix of the model of Figure 1.1. 6-P-G-1,5-L denotes 6-P-glucono-1,5; S-7-P denotes sedoheptulose-7-P and G-3-P denotes glyceraldehyde-3-P. Reaction ρ_6 requires two molecules of xylulose-5-P to produce a sedoheptulose-7-P molecule and a glyceraldehyde-3-P molecule.

2.2 Steady state metabolic flux analysis

The methods we propose for metabolic flux analysis belong to the *stoichiometric* paradigm of metabolic modelling [KS03]. In the stoichiometric model the total sum of cellular reactions of the same kind are lumped together to provide a comprehensive model of the metabolism [SAN98]. For every lumped reaction its substrates and products as well as the molar ratios in which substrates are consumed and products produced by the reactions are specified. The stoichiometric model can be represented as a bipartite graph that is composed of metabolite and reaction nodes (see Figure 1.1 for an example). It is useful to describe the stoichiometry of an organism as a *stoichiometric matrix* that has a column for each reaction and a row for each metabolite. Coefficients of the matrix then define the molar ratios for consumption and production of metabolites in reactions. Formally, the stoichiometric matrix A corresponding with a metabolic network G is a matrix of m rows and n columns. The coefficients $A(i, j)$ are equal to the number α_{ij} of metabolite molecules M_i produced or consumed in a single reaction event of ρ_j. Table 2.1 presents the stoichiometric matrix of the metabolic network in Figure 1.1.

A major simplification made in the stoichiometric modelling paradigm is to leave the reaction kinetics, that is, dynamics that describe the reaction mechanisms, regulation and enzyme properties [HS96, Hei05, MK98, MMB03] out of the model. This seriously limits the applicability of the

modelling paradigm in the study of the regulation and the dynamic behaviour of metabolism. On the other hand, the stoichiometry of central metabolism is relatively well understood for many organisms, while the detailed reaction mechanisms and enzyme properties are not – at least not in the systemic scale and the level required for quantitative modelling [SAN98, PSP03, WT04]. Thus by leaving the kinetics out, stoichiometric models can be based on more reliable information, with the cost of giving up on the detailed dynamic modelling of metabolism and its regulation.

Stoichiometric models have proven to be useful in many tasks of metabolic modelling [KS03]. In metabolic pathway analysis, functional, biochemically meaningful pathways are identified from stoichiometric models [PPW$^+$03, SSPH99, SFD00]. Maximal yields of end products of a metabolism, that is, the ratio of the amount of specific targets produced and external substrates consumed, can be computed from the stoichiometric models [SKWP02]. Furthermore, based on the stoichiometry, it is possible to design genetic modifications to an organism to improve the yields of specific target metabolites [BM03, PBM04] and to approximate the robustness of the metabolism to genetic mutations and to environmental changes [SKB$^+$02].

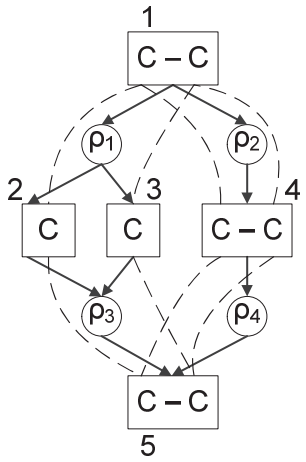
In metabolic flux analysis, we usually assume that rates v_j of reactions $\rho_j \in \mathcal{R}$ and the sizes of metabolite pools stay constant over time, that is, the metabolism of a cell is assumed to be in *steady state*. In such a state the *metabolite balance*, or *mass balance*

$$\sum_{j=1}^{n} \alpha_{ij} v_j = \beta_i \tag{2.1}$$

holds for each metabolite M_i. Here, β_i is the measured external inflow ($\beta_i < 0$) or external outflow ($\beta_i > 0$) of metabolite M_i. From balance equations (2.1) defined for every metabolite M_i we will obtain a metabolite balancing, or stoichiometric equation system

$$A\mathbf{v} = \beta, \tag{2.2}$$

constraining the fluxes $\mathbf{v}$. For simple tree-like network topologies that do not contain cycles, bidirectional reactions or alternative routes between source and target metabolites, (2.2) is fully determined linear system and fluxes $\mathbf{v}$ can be solved from it with standard matrix pseudoinverse. However, for realistic metabolic networks (2.2) is underdetermined. By analyzing the null space of matrix A [KS02], it is possible to solve from the underdetermined (2.2) some fluxes whose values are the same in every feasible



	ρ_1	ρ_2	ρ_3	ρ_4
M_1	-1	-1	0	0
M_2	1	0	-1	0
M_3	1	0	-1	0
M_4	0	1	0	-1
M_5	0	0	1	1

Figure 2.3: Two competing pathways from metabolite M_1 to M_5 and the corresponding underdetermined stoichiometric matrix.

flux distribution, but in general case solutions to (2.2) contain $(n-rank(A))$ free fluxes, whose values need to be fixed by some other means. Figure 2.3 depicts a small metabolic network with two alternative routes from M_1 to M_5 and the corresponding stoichiometric matrix. The sum of columns 1 and 3 corresponding pathway (ρ_1, ρ_3) equals the sum of columns 2 and 4 corresponding pathway (ρ_2, ρ_4). Thus the linear equation system defined by the stoichiometric matrix is underdetermined, even if the intake of M_1 and the output of M_5 can be measured.

One possibility to estimate the steady-state fluxes that are not fully constrained by (2.2) is to make an additional assumption that the metabolism of a modelled organism has an objective, such as *optimal growth*, that it tries to fulfill in the given conditions. In *flux balance analysis* [VP94, BST97, ECP02] this objective is coded as a linear function of fluxes. The task is then to maximimise the value of an objective function in the feasible space spanned by the stoichiometry and the constraints $v_i^{min}$and v_i^{max} stating minimum and maximum allowable values for each flux v_i. Thus we obtain flux distribution $\mathbf{v}$ from a linear programming [Mar01] problem of the following form:

$$\begin{aligned} &\max_{\mathbf{v}} \sum_i c_i v_i \\ \text{s.t.} \quad &A\mathbf{v} = \beta \\ &v_i^{min} \leq v_i \leq v_i^{max} \ \forall v_i \in \mathbf{v}. \end{aligned} \tag{2.3}$$

It has been empirically shown that in certain conditions, the biomass

yield, i.e. the growth, of bacteria *E. coli* is indeed optimal within the constraints posed by the stoichiometry [EIP01]. The flux balance framework has also been successfully applied to predict the lethality of gene deletions by computing the optimal growth rates for networks without reactions catalyzed by genes whose lethality is to be tested [EP00]. Recently, flux balance analysis was used to predict gene interaction networks by computing growth optimal fluxes for all single and double knockouts of 890 metabolic genes of *Saccharomyces cerevisiae* [SDCK05].

However, flux balance analysis alone is not the ultimate tool for metabolic flux analysis [ECP02, FS05, MH03]: First, the behaviour of cells is not necessarily stoichiometrically optimal. Second, the true objectives might be unknown for every condition or after every genetic modification. Third, in general the flux vector maximizing (2.3) is not unique. More information is thus required to obtain knowledge about the fluxes in given conditions.

2.3 Isotopic labelling experiments

Currently, the most accurate estimates of the fluxes in a metabolic network are gained when the stoichiometric information is augmented with information obtained from isotopic labelling experiments. In an isotopic labelling experiment a cell population is fed with labelled nutrients, such as glucose containing ^{13}C atoms. Labels are then transferred by chemical reactions to other metabolites where they induce different isotopomer distributions depending on the rates and the carbon mappings of reactions in the network.

If in addition to the reaction rates, isotopomer distributions of metabolites remain constant, the metabolic network is in an *isotopomeric steady state*. In such a state, the rate of production and consumption of each isotopomer $M_i(b)$ of each metabolite M_i satisfies the *isotopomer balance* (cf. (2.1))

$$\sum_{j=1}^{n} \alpha_{ij} v_j P_{M_{ij}}(b) = \beta_i P_{M_{i0}}(b) \tag{2.4}$$

for any $b \in \{0,1\}^{|M_i|}$.

In (2.4), the isotopomer distributions of the outflow subpools of M_{ij} ($\alpha_{ij} < 0$) are always identical to the distribution of the whole mixed metabolite pool M_i as we assume that reactions uniformly sample their reactant pools (see Section 3.1). If, however, the pathways leading to a junction metabolite—a metabolite with more than one producer—manipulate the carbons of the metabolite differently, then the isotopomer distributions

of the inflows ($\alpha_{ij} > 0$) often have differences. Because of these differences equations of (2.4) can be linearly independent, and constrain the fluxes more than mass balance equations (2.1) alone. If, for example, in Figure 2.3 $P(M_1(00)) = 0.9$ and $P(M_1(11)) = 0.1$, subpools M_{53} and M_{54} will get different isotopomer distributions, because pathway (ρ_1, ρ_3) cleaves the carbon chain and thus mixes the fragments of unlabelled and totally labelled metabolite molecules, while pathway (ρ_2, ρ_4) transports molecules intact from M_1 to M_5. For example, $P(M_{53}(00)) = 0.9 \cdot 0.9 = 0.81$ and $P(M_{54}(00)) = 0.9$. If we add a constraint

$$\sum_{j=1}^{n} \alpha_{5j} v_j P_{M_{5j}}(00) = \beta_5 P_{M_{50}}(00) \Leftrightarrow$$
$$v_3 \cdot 0.81 + v_4 \cdot 0.9 = \beta_5 P_{M_{50}}(00)$$

to the stoichiometric system of Figure 2.3, and are able to measure $P_{M_{50}}(00)$ and inflow of M_1 or the outflow M_5, the system will be fully determined and all fluxes can be solved.

Thus, by measuring the isotopomer distributions from metabolites, information about the fluxes of competing pathways, cycles and bidirectional reactions can be obtained.

2.4 Measurement technologies

Today, isotopomer distributions can be measured with two basic technologies, by nucleic magnetic resonance spectroscopy (NMR) [MdGW$^+$96, SGH$^+$99] or mass spectrometry (MS) [CN00, DS00, FNS04, WH99]. In this section we shortly describe the type of constraints these instruments can measure to isotopomer distributions. The emphasis of the introduction is in MS, which is more central to this thesis.

2.4.1 Nucleic magnetic resonance spectroscopy

In a widely applied 2D [^{13}C ,^1H] COSY (HSQC) technique of NMR, ^{13}C atoms coupled to an observed ^{13}C atom through one-bond couplings or long-range couplings give rise to characteristic signal fine structure in a NMR spectrum [Szy95] (see Figure 2.4 for an example). By analyzing the relative intensities of the signal fine structures from different combinations of the coupled ^{13}C atoms, constraints to the isotopomer distribution of the metabolite measured can be inferred [SGH$^+$99, vWSVH01].

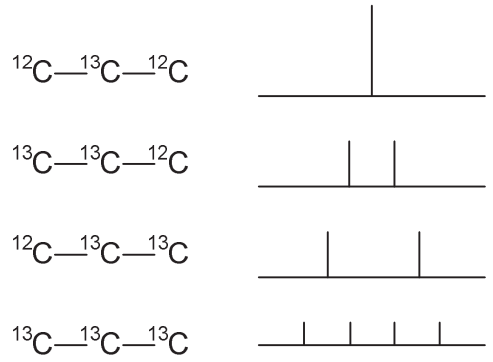


Figure 2.4: Different combinations of ^{13}C and ^{12}C atoms that are coupled to an observed ^{13}C atom (in the middle) give rise to characteristic signal fine structures in NMR spectrum. The heights of the peaks (y-axis) are proportional to the relative intensity of the corresponding isotopomer.

For example, for metabolite M_i with carbon chain of length three, the following constraints to $D(M_i)$ can be inferred:

$$\frac{P_{M_i}(b'1b'')}{\sum_{b_1,b_3\in\{0,1\}} P_{M_i}(b_1 1 b_3)} = d_{i,(b'1b'')} \tag{2.5}$$

for each label combination $M(b'1b'')$, where $d_{i,(b'1b'')}$ is the measured relative intensity of a peak in an NMR spectrum corresponding isotopomer $(b'1b'')$. Using 2D $[^{13}C\,,^{1}H]$ COSY NMR measurements different label combinations can be observed around ^{13}C atoms bound to at least one hydrogen atom. Thus, neither the ^{12}C atoms and their adjacent carbons nor the labelling status of the carbons adjacent to the carboxyl group (COOH) carbon can be observed. Thus, the complete isotopomer distributions cannot be uncovered in general. (With small, isolated metabolites, this problem can be circumvented by applying ^{1}H heteronuclear spin difference NMR spectroscopy [dGMM^{+}00].) Furthermore, the sensitivity usually limits the applicability of NMR spectroscopy to detection of proteinogenic amino acids abundant in the cell biomass while the isotopomer distributions of the internal primary metabolites remain undetectable due to their low concentrations in cells.

2.4.2 Mass spectrometry

Mass spectrometer (MS) measures the abundances of molecules with different masses in a sample with very high precision [MZSL98]. There exist

many different mass spectrometry techniques that all contain the same basic steps. First molecules are ionized by an *ion source*. Ionization gives molecules an electric charge so they can be moved with electronic fields. Then the *mass analyzer* separates ions according to their mass-to-charge ratio (m/z)[1]. In the third phase of MS measurement the *detector* records the charge induced or the current produced when an ion passes by or hits a surface from which the number of ions with specific m/z value can be deduced.

In tandem MS (MS-MS) [McL80] two or more mass analyzers are used in succession to fragment molecules and to also measure the abundances of the fragments with different weights. The fragmentation of molecular ions can be achieved by many techniques. In a common collision-induced dissociation (CID) method [Mar98] metabolite molecules are collided with neutral gas which results in bond breakage and the fragmentation of a molecular ion. For the purposes of this thesis, two different modes of fragmentation are distinguished. In *full scanning mode* all mass isotopomers of a metabolite are simultaneously fragmented. In *daughter ion scanning* mode every mass isotopomer of the metabolite can be separately fragmented and the mass isotopomer distributions of fragments measured. In general, this separation produces more constraints to the isotopomer distribution, as shown in Publication IV and Publication V. It also affects the computation of constraints to the isotopomer distribution from MS-MS data (see Chapter 5 for more details). Figure 2.5 depicts a daughter ion spectrum of ^{13}C labelled alanine.

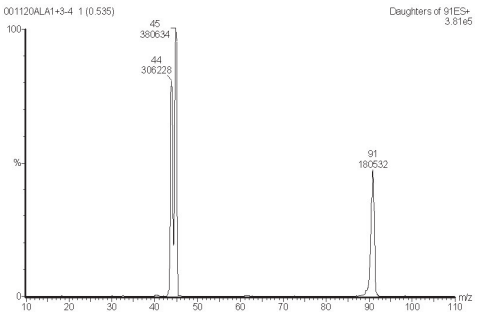


Figure 2.5: Daughter ion spectrum of ^{13}C labelled alanine (fragmentation at m/z 91 Da, figure from Publication IV).

[1]Small molecules such as metabolites are usually single charged. Thus the mass-to-charge ratio can be thought to be equal to the mass of an ion.

Before entering MS, metabolite molecules are usually separated by their chemical properties (liquid chromatography, LC), their boiling points (gas chromatography, GC) or by their mobility in a capillary (capillary electrophoresis, CE). Thus in ^{13}C metabolic flux analysis, MS can be used to measure the mass isotopomer distributions of a metabolite $\mathcal{M}_i$ [CN99, FNS04], that is, constraints

$$P_{\mathcal{M}_i}(+k) = d_{i,k} \tag{2.6}$$

to $D(M_i)$, where $d_{i,k}$ is the relative intensity of a peak in MS spectrum corresponding $\mathcal{M}_i(+k)$. More information about the isotopomer distributions can be acquired by applying MS-MS to obtain analogous constraints

$$P_{\mathcal{M}_i|\mathcal{F}_j}(+k) = d_{i,j,k} \tag{2.7}$$

to isotopomer distributions of fragments $\mathcal{F}_j$ emerging in MS-MS. Chapter 5 of this thesis introduces methods to compute constraints to the isotopomer distribution $D(M_i)$ of the carbon chain of $\mathcal{M}_i$ from (2.6) and (2.7). The sensitivity of MS-MS methods is generally better than NMR's, but still some metabolites cannot be reliably analyzed because of the low abundance or chemical properties of the compound. The amount of independent constraints obtained to isotopomer distribution depends on the fragmentation pattern of a metabolite in MS-MS. In general, full isotopomer distributions are not uncovered.

2.5 General model for measurement data

Above we learned that neither NMR nor MS-MS can measure full isotopomer distributions $D(M_i)$ for each metabolite M_i in the network. Thus (2.4) cannot be directly applied to solve the fluxes. Instead, both technologies measure linear constraints

$$\sum_b s_{b,i,h} P_{M_i}(b) = d_{ih}, \tag{2.8}$$

to $D(M_i)$, where d_{ih} is the measured relative abundance of the specified linear combination of isotopomers. The coefficients $s_{b,i,h} \in \mathbb{R}$ depend on the measurement technique and the metabolite. We apply this simple, yet general model of isotopomer measurement data to develop computational methods that can simultaneously make use of NMR and MS-MS measurements — or linear constraints to isotopomer distributions obtained by some other means.

Finally, a geometrical interpretation of linear constraints to the isotopomer distribution will be useful later in the thesis. Isotopomer distribution $D(M)$ defines a point in the *isotopomer space* $\mathcal{I}_M$ spanned by the standard vectors $\mathbf{i_b} \in \{0,1\}^{2^{|M|}}$ that contain 0's in all other components except in the b'th location. More generally, a set of linear constraints to the isotopomer distribution $D(M)$, such as mass isotopomer distribution $D(M)^m$ or general measurement constraints (2.8), defines a linear subspace of $\mathcal{I}_M$.

Chapter 3

^{13}C metabolic flux analysis

This chapter introduces the basic assumption behind ^{13}C metabolic flux analysis and define the problem of ^{13}C metabolic flux estimation. Also, existing computational methods for ^{13}C metabolic flux analysis are reviewed.

3.1 Modelling assumptions

^{13}C metabolic flux analysis is commonly based on a few key assumptions about the modelled metabolism (cf. [Wie02]).

1. A cell population has reached isotopomeric steady state before the isotopomer measurements are conducted.

2. The state of individual cells in the population is not too different from the population average.

3. The model of metabolism is complete, that is, all reactions with nonzero flux of an organism that produce or consume the metabolites in the model are present and the carbon mappings are correct for each reaction in the model.

4. Metabolites and enzymes are fully mixed in the cell compartments.

5. Reactions sample substrate pools uniformly, thus different isotopomers are consumed in the proportion of their abundances.

6. Reactions in the model are simple, that is, they do not contain hidden intermediate steps where substrate molecules are drawn from mixed pools.

Together, these assumptions justify the writing of flux balance equations (2.4). They also facilitate the propagation of measurement data in the metabolic network with methods described later in the thesis. According to assumption (1), isotopomer distributions of metabolites stay constant over time (for techniques to avoid this assumption, see [NW06]). Assumption (2) makes it possible to state that the flux estimations based on the measurements from a cell population hold reasonably well for individual cells, too. Assumption (4) is the basic assumption behind the lumping of reactions in stoichiometric modelling. It implies a "Markov property" of a metabolic network: reactions that have the same substrates sample them from a common pool, disregarding the history of specific substrate molecules. Assumption (5) states that extra labels in metabolite molecules have no effect on the use of molecules as substrates. Assumption (6) requires that all reaction steps that affect on labelling patterns of metabolite molecules are explicitly modelled.

3.2 Problem of ^{13}C flux estimation

As mentioned in Section 2.4, current measurement technologies can only measure linear constraints to isotopomer distributions of some metabolites in the network. Thus isotopomer balances (2.4) cannot be directly applied to uncover the fluxes of a metabolic network. A general formulation for the problem of ^{13}C flux estimation that allows missing measurement data models a metabolic network as the nonlinear system of fluxes and isotopomer distributions. In this system, the stoichiometry of the network and measured external flows and constraints to the isotopomer distributions of metabolites in the network need to be described. Furthermore, the carbon mappings of reactions and the random, unbiased sampling of different isotopomers of substrates by the reactions need to be modelled.

For the modelling of the random sampling of the substrate isotopomers and the carbon mappings, two technical concepts need to be introduced. In reaction ρ_j, the carbon mapping λ_j between substrate and product metabolites of ρ_j also defines a carbon mapping $\lambda_j^{k,l} : M|F_k \rightarrow M'|F_l$, where M is a substrate and M' a product of ρ_j. Furthermore, $\lambda_j^{k,l}$ induces an *isotopomer mapping* $\iota_j^{k,l} : \mathcal{I}_{M|F_k} \rightarrow \mathcal{I}_{M'|F_l}$ of the isotopomers of substrate fragment $M|F_k$ to the isotopomers of product fragment $M'|F_l$. In $\iota_j^{k,l}$ isotopomer $F_k(b)$ is mapped to $F_l(b')$ if $\lambda_j(b) = b'$. In the example of Figure 3.1, substrates M_1 and M_2 are mapped to fragments $M_3|F_2$ and $M_3|F_1$ of product M_3. In the example, $\iota^{M_1,M_3|F_2}(01) = 10$ and $\iota^{M_2,M_3|F_1}(0) = 0$. Thus, $P_{M_3}(010) = P_{M_2}(0) \cdot P_{M_1}(01)$. An *isotopomer*

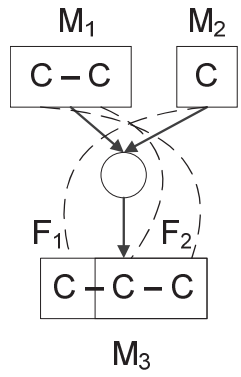


Figure 3.1: An example of isotopomer mappings.

mapping matrix $IMM_j^{k,l}$ [SCNV97b] of fragments F_k and F_l is a square binary matrix with $2^{|F_k|}$ rows and columns. Coefficient $IMM_j^{k,l}(b, b') = 1$ if $\iota_j^{k,l}(b) = b'$, otherwise $IMM_j^{k,l}(b, b') = 0$. Second, let $M_m|G$ and $M_n|H$ be substrate fragments mapped to M_{ij} by reaction ρ_j, $|M_{ij}| = |G| + |H|$. Now, because of the random and unbiased sampling of different substrate isotopomers by the reactions, the abundance of an isotopomer $M_{ij}(b)$ can be computed by multiplying the abundances of isotopomers $G(b')$ and $H(b'')$ that make up $M_{ij}(b)$, after appropriate isotopomer mappings are first applied to substrate fragments. That is,

$$P_{M_{ij}}(b) = P_{M_m|G(b')} \cdot P_{M_n|H(b'')}|\iota_j(G(b') \cup H(b'')) = M(b). \tag{3.1}$$

More generally, the isotopomer distribution of the product metabolite is computed by taking component-wise Kronecker product $\otimes$ of the isotopomer distributions of the substrate fragments [1].

[1]The result of $m \otimes n$ is a matrix formed from all possible products of the elements of m with those of n. If m is l-by-k and n is p-by-q, then $m \otimes n$ is lp-by-kq. The elements are arranged in the following order:

$$\begin{matrix} [m(1,1) * n & m(1,2) * n & \dots & m(1,k) * n \\ & & \dots & \\ m(l,1) * n & m(l,2) * n & \dots & m(l,k) * n]. \end{matrix}$$

Problem 1 (^{13}C Flux Estimation Problem)**.** Solve the flux distribution $\mathbf{v} = (v_1, \ldots, v_n)$ from the equation system

$$\sum_{j=1}^{n} \alpha_{ij} v_j P_{M_{ij}}(b) = \beta_i P_{M_{i0}}(b) \qquad \forall M_i \tag{3.2}$$

$$M_{ij} = M_i \qquad \forall M_{ij} : \alpha_{ij} < 0 \tag{3.3}$$

$$D(M_{ij}) = \bigotimes_{\alpha_{kj}<0} IMM_j^{k,i} D(M_k|F_{ij}) \qquad \forall M_{ij} : \alpha_{ij} > 0 \tag{3.4}$$

$$\sum_{b} s_{b,i,h} P_{M_i}(b) = d_{i,h} \qquad \forall M_i, h \tag{3.5}$$

where $d_{i,h}$'s denote measured isotopomer constraints, $M_k|F_{ij}$ denotes a fragment of metabolite M_k that is mapped to metabolite M_i by reaction ρ_j and $\bigotimes$ denotes a series of consecutive pairwise Kronecker products $\otimes$.

Bilinear equation (3.2) models the isotopomeric steady state. Equation (3.3) states that the isotopomer distributions of outflow subpools are identical to the isotopomer distribution of the mixed pool of a metabolite. Nonlinear equation (3.4) states that the isotopomer distribution of a product metabolite pool of reaction ρ_j is the product of the isotopomer distributions of the substrates of ρ_j. Equation (3.5) integrates isotopomer measurements to the model.

3.3 Existing approaches of ^{13}C metabolic flux analysis

At high level, there exist two approaches to tackle Problem (1). These approaches are introduced next.

3.3.1 Optimization methods for ^{13}C metabolic flux analysis

A popular approach for ^{13}C metabolic flux analysis formulates Problem (1) as a constrained least-squares minimization problem, where the difference between the observed and simulated isotopomer measurements is minimized [AKS06, SCNV97a, WdG97, WMdG01, YWH04]. Let $\mathbf{x} = (D(M_i))_i$ denote a vector collecting the relative abundances of every isotopomer of every metabolite in the network. It has been shown [WW01] that, for any given steady-state flux distribution $\mathbf{v}$ and any practically relevant metabolic network, there exists exactly one isotopomer steady state $\mathbf{x}^{\mathbf{v}}$. Let

$\mathbf{x}^{inp}$ denote a vector of relative abundances of the isotopomers of external substrates and let $\mathbf{x}^{obs}$ denote a vector of measured constraints to the isotopomer distributions of the metabolites. Let $\mathbf{y}^{obs} = [\mathbf{x}^{obs}\beta]$ denote a vector where measured inflows and outflows for each metabolite are concatenated to $\mathbf{x}^{obs}$, let $\beta^{\mathbf{v}}$ denote inflows and outflows defined by $\mathbf{v}$ and let $\mathbf{y}^{\mathbf{v}} = [\mathbf{x}^{\mathbf{v}}\beta^{\mathbf{v}}]$. Then a solution to Problem (1) is found by solving a least-squares minimization problem

$$\min_{\mathbf{v}} \|\mathbf{y}^{obs} - \mathbf{y}^{\mathbf{v}}\|_2^2. \tag{3.6}$$

As isotopomer distributions $\mathbf{x}^{\mathbf{v}}$ depend from fluxes $\mathbf{v}$ in nonlinear fashion, the minimization problem (3.6) is typically solved with iterative methods. At each step of the iteration, fluxes are fixed to some candidate values and isotopomer distributions of metabolites are computed from the network model. By a proper transformation of isotopomer data, $\mathbf{x}^{\mathbf{v}}$ can be computed analytically from known $\mathbf{v}$ and $\mathbf{x}^{obs}$, by solving a sequence of linear equation systems [WMI+99]. If the least squares difference (3.6) is small enough, candidate fluxes are returned as a result. Otherwise some optimization method, such as a gradient-based method [WMdG01] or an evolutionary or simulated annealing algorithm [DBS01, SNV99] is used to select new candidate fluxes that are likely to produce smaller difference and the iteration is continued. We call methods for ^{13}C metabolic flux analysis that apply this general strategy *optimization methods*. More detailed descriptions of the optimization methods are available in the introductory texts [WMdG01] and [Wie02].

The optimization approach for ^{13}C metabolic flux analysis is applicable to all network topologies, it can give separate estimations for both directions of reversible fluxes and it can easily utilize all the measurement data available. There also exists a widely used implementation of an optimization framework, 13C-FLUX [WMdG01]. A faster version of the general optimization framework is also tailored for NMR measurements and uniform substrate labellings [vWHV02].

As a drawback the optimization framework shares the problems of general nonlinear optimization methods. First, it is hard to make sure that an optimization process has converged to a global instead of a local minimum [GZG+05]. Second, if measurement data does not fully determine the flux distribution, optimization methods will merely sample the solution space, and cannot return the set of all feasible solutions. By performing an a priori identifiability analysis [IW03, vWHVG01] and analyzing the sensitivity of a point solution [MWKdG99] or computing the confidence intervals for estimated fluxes [AKS06] it is possible to examine the uniqueness

of the solution. The mathematical complexity of these statistical analysis methods is quite high.

3.3.2 Direct methods for ^{13}C metabolic flux analysis

Another, "direct" approach [2] for ^{13}C metabolic flux analysis is based on the general idea that some ratios of fluxes producing the same junction metabolite can be inferred from isotopomer data without iterative fitting of the fluxes to isotopomer measurements. Together with mass balances (2.1) defined by the stoichiometry of the network, these flux ratios can give enough information about the fluxes and nonlinear optimization can be avoided.

The ratios of fluxes producing metabolite M_i are derived from isotopomer balances (2.4) constraining the fluxes around M_i. According to (2.4), the mass balance (2.1) holds for each isotopomer separately. Thus the similar balance holds for any linear combination of isotopomers, too. Formally,

$$\sum_{j=1}^{n} \alpha_{ij} v_j d_{ijh} = \beta_i d_{i0h}, \tag{3.7}$$

for all metabolites M_i, where d_{ijh} is a linear combination (cf. (2.8)) from

$$\sum_{b} s_{b,i,h} P_{M_{ij}}(b) = d_{ijh}, \tag{3.8}$$

of isotopomer abundances known for subpool M_{ij}. We note that coefficient $s_{b,i,h}$ is the same for all subpools M_{ij}, that is, all d_{ijh}'s define points from the same subspace of isotopomer space $\mathcal{I}_M$. If enough independent equations of the form (3.7) can be written, the ratios of fluxes producing M_i can be solved from the resulting linear equation system. In the best case, all flux ratios are uncovered. In that case, all information about the flux distribution obtainable from isotopomer measurements is also uncovered. Thus, direct methods for ^{13}C metabolic flux analysis that avoid difficulties of nonlinear optimization can be as powerful as optimization methods for estimating steady state fluxes. On the other hand, in order to write (3.7) to junction metabolite M_i, we need to know (3.8) for identical linear combination $s_{b,i,h}$ of isotopomers of M_i for each subpool M_{ij}. Thus a direct framework might require more measurement data than an iterative framework to estimate the complete flux distribution. The central question for

[2] Here, "direct" refers only to the general strategy to obtain and utilize flux ratios directly, instead of iterative fitting of fluxes to the measurements. Methods labelled here as "direct" may also use iterative optimization methods in some of their stages.

direct methods of ^{13}C metabolic flux analysis – and also for this thesis – then is: *How to obtain enough common isotopomer constraints (3.8) for all subpools of junction metabolites from incomplete measurements to uncover the fluxes of interest?*

METAFoR analysis

The origins of the direct approach for ^{13}C metabolic flux analysis lie in METAFoR (Metabolic Flux Ratio) analysis [Szy95, SGH$^+$99]. In MeTAFoR analysis, the ratios of fluxes producing a junction metabolite are studied locally, in separation from the rest of the network. In traditional METAFoR analysis, a uniform labelling of external substrates that consists of a mixture of unlabelled and fully labelled molecules is applied and NMR is used as a measurement technique. The flux ratios are then computed for junction metabolites M_i produced by two competing pathways $\mathcal{P}_1$ and $\mathcal{P}_2$ of which $\mathcal{P}_1$ keeps carbon–carbon bond (c_k, c_l) of M_i intact all the way from an external substrate while in $\mathcal{P}_2$, bond (c_k, c_l) is created by a reaction of the pathway. Thus the ratio of intact fragments (c_k, c_l) corresponds to the relative flux through $\mathcal{P}_1$. The constraints (2.5) measurable by NMR fit well for uncovering this ratio of intact fragments. Intuitively, if the label of c_k differs from the label of c_l, bond (c_k, c_l) is biosynthetically created and the metabolite molecule produced by $\mathcal{P}_2$, while if the labels of c_k and c_l are the same, it is more probable that bond (c_k, c_l) has stayed intact and a molecule is produced by $\mathcal{P}_1$. With careful modelling [Szy95] that takes into account that also $\mathcal{P}_2$ can produce molecules that have the same labels in c_k and c_l and that unlabelled external substrates contain ^{13}C atoms in their natural abundance, the ratios of fluxes through $\mathcal{P}_1$ and $\mathcal{P}_2$ can be accurately computed. For example, let $P_n(0)$ $P_n(1)$ denote the relative natural abundance of ^{12}C and ^{13}C atoms, respectively. Let $v_1 \in \mathcal{P}_1$ and $v_2 \in \mathcal{P}_2$ be fluxes of reactions producing $M_i = (c_k, c_l, *)$ ($*$ denotes some unspecified chain of carbon atoms) and let v_3 be the flux of the only reaction consuming M_i. Let $P_S(^{13}C)$ be the fraction of totally labelled and $P_S(^{12}C) = 1 - P_S(^{13}C)$ the fraction of unlabelled substrate molecules. Because of the uniform labelling of substrates, the relative abundance $P_c(1)$ of labels in every carbon c in the network is the same.

An NMR measurement observing carbon $c_k \in M_i$ that is connected to

only one other carbon gives isotopomer constraints (cf. 2.5)

$$\frac{P_{M_i|(c_k,c_l)}(10)}{N} = d_{i1}$$
$$\frac{P_{M_i|(c_k,c_l)}(11)}{N} = d_{i2},$$

where $N = P_{M_i|(c_k,c_l)}(10) + P_{M_i|(c_k,c_l)}(11) = P_c(1)$. Let $\mathbf{d}_i = [d_{i1}d_{i2}]$. From the properties of pathways $\mathcal{P}_1$ and $\mathcal{P}_2$ we can deduce similar isotopomer constraints for subpools $M_{i1}|(c_j, c_k)$ and $M_{i2}|(c_j, c_k)$:

$$d_{i11} = \frac{P_{M_{i1}|(c_k,c_l)}(10)}{N} = \frac{P_S(^{12}C)P_n(1)P_n(0)}{P_c(1)}$$
$$d_{i12} = \frac{P_{M_{i1}|(c_k,c_l)}(11)}{N} = \frac{P_S(^{13}C) + P_S(^{12}C)P_n(1)^2}{P_c(1)}$$
$$d_{i21} = \frac{P_{M_{i2}|(c_k,c_l)}(10)}{N} = \frac{P_c(1)P_c(0)}{P_c(1)}$$
$$d_{i22} = \frac{P_{M_{i2}|(c_k,c_l)}(11)}{N} = \frac{P_c(1)P_c(1)}{P_c(1)}$$

Thus we can construct generalized isotopomer balances (cf. (3.7))

$$\alpha_{i1}\mathbf{d}_{i1}v_1 + \alpha_{i2}\mathbf{d}_{i2}v_2 - \alpha_{i3}\mathbf{d}_i v_3 = 0$$
$$\alpha_{i1}v_1 + \alpha_{i2}v_2 - \alpha_{i3}v_3 = 0$$

bounding the ratios of fluxes v_1 and v_2. Analogous equations can be written also for central carbons of metabolites, such as $c_l \in M_i$ [Szy95].

As the flux ratios are computed for every junction separately in METAFoR, global flux distribution is not obtained. In traditional METAFoR, flux ratios are computed for the junctions that are produced by competing pathways of which some cleave and recombine the external substrates while some do not. More generally, the METAFoR approach can be applied, if the fraction of fragments that have stayed intact from external substrates can be inferred from the measurements for each subpool of a junction metabolite (see e.g.[MFC+01]). However, for large models containing bidirectional reactions and cycles these junctions are nontrivial to find manually. The limited sensitivity of NMR usually restricts the set of measurable metabolites to amino acids abundant in the cell. Also, uniform labelling of external substrates is not optimal for solving all the fluxes in metabolic networks [FS03]. Because of these properties, METAFoR from NMR measurements is not always able to uncover all the flux ratios of interest.

^{13}C constrained flux balancing

Local flux ratios of METAFoR (or constraints to them) can be easily combined with stoichiometric constraints ([SHB+97] and Publication I). With the above notation, it is enough to augment a stoichiometric equation system containing mass balances (2.1) for each metabolite with generalized isotopomomer balances (3.7) constraining the flux ratios of (some) junctions. This kind of approaches are sometimes called the methods of ^{13}C constrained flux balancing [SHB+97]. Recently, a method for ^{13}C metabolic flux analysis where (2.1)'s are augmented with flux ratios derived from gas chromatography mass spectrometer (GS-MS) data was presented [FS03, FNS04]. Also, a software called *FiatFlux* implementing the method was provided [FNS05]. In FiatFlux, mass isotopomer distributions of intermediate precursor metabolites are propagated from the mass isotopomer distributions of the fragments of amino acids subjected to GC-MS. After the propagation of mass isotopomer data the flux ratios are computed by applying (3.7) where each d_{ijk} corresponds to the abundance of the same mass isotopomer $+k$ of subpool M_{ij}, that is,

$$P_{M_{ij}}(+k) = d_{ijk},$$

for each relevant mass isotopomer $M_{ij}(+k)$. Finally, the computed flux ratios are combined with (2.1)'s for each metabolite. The authors solve the flux distribution from the resulting equation system by formulating the problem as a nonlinear least squares optimization problem where the computed flux ratios are weighted with their experimental variances and their magnitude [FNS04]. Formally, let $\mathcal{R}^i = \rho^i_1, \ldots, \rho^i_k$ denote reactions producing metabolite M_i and let $R_{mb} = A\mathbf{v} - \beta$ be a residual error of the equation system (2.2). Let f_i denote the computed ratio of the sums of fluxes of two subsets $\mathcal{R}^i_1$ and $\mathcal{R}^i_2$, $\mathcal{R}^i_1 \bigcup \mathcal{R}^i_2 = \mathcal{R}^i$ and let residual R_i be defined as

$$R_i = f_i \sum_{\rho_p \in \mathcal{R}^i_2} \alpha_{ip} v_p - \sum_{\rho_q \in \mathcal{R}^i_1} \alpha_{iq} v_q. \tag{3.9}$$

Then in FiatFlux, the fluxes are obtained by solving a problem

$$\min_{\mathbf{v}} \frac{(R_{mb})^2}{(\sigma_{mb})^2} + \sum_i \frac{(R_i)^2}{(\sigma_i)^2 (\frac{\partial R_i}{\partial f_i})^2}, \tag{3.10}$$

where experimental variances σ are either estimated from the measurements or assigned a priori (for handling of upper bounds to the flux ratios, see [FNS04]).

In FiatFlux, mass isotopomer distributions of internal metabolites are derived from the measured mass isotopomer distributions of amino acid fragments by constructing a nonlinear equation system constraining the components of precursor mass isotopomer distributions [FS03]. If, for example, a measured amino acid AA originates from precursors M_1 and M_2, the mass isotopomer distributions $D(M_1)^m$ and $D(M_2)^m$ are obtained by constructing and solving a bilinear equation system that has an equation

$$P_{AA}(+k) = \sum_{i+j=k} P_{M_1}(+i)P_{M_2}(+j) \tag{3.11}$$

for each mass isotopomer $P_{AA}(+k)$ of AA. As a general propagation method, this technique is powerful and applicable with all labellings of external nutrients. However, the technique is not free from the problems of multiple optimal solutions and convergence to a local optima. In the worst case, computed flux ratios can be spurious due to the convergence to a wrong optimum during the computation of mass isotopomer distributions of precursors. The possibility of selecting the wrong optimum can be illustrated with an (pathologic) example where two carbon amino acid AA is produced by combining two one-carbon precursors M_1 and M_2. Let us assume that all molecules of AA have exactly one ^{13}C label in their backbone, thus $P_{AA}(+1) = 1$. Now, either $P_{M_1}(+1) = 1$ or $P_{M_2}(+1) = 1$. Both alternatives are equally good solutions of (3.11), but only one of them can be correct. If the wrong solution is selected, the flux ratios that are computed based on $D(M_1)$ and $D(M_2)$ can also be wrong (for further discussion, see [WDW04]).

Starting from the next chapter of the thesis, we present a framework for ^{13}C constrained flux balancing. Like FiatFlux, our method consists of three main steps: 1) the propagation of measurement information in a metabolic network 2) augmentation of mass balances (2.1) with generalized isotopomer balances (3.7) and 3) solving the resulting equation system. However, in the propagation step we apply only techniques that are "safe", that is, we only propagate isotopomer constraints if we can be sure that the propagated constraints must hold, if our modelling assumptions are correct. Also, our method is not tied to any specific type of measurement technology but can simultaneously utilize all isotopomer data that can be described as linear constraints to the isotopomer distribution. For example, measurements produced by common NMR, MS and tandem MS techniques can be simultaneously utilized. Thus, our method can be seen as a generalization of previous methods for direct ^{13}C metabolic flux analysis that are suitable only for a certain kind of data, such as mass isotopomer or NMR data, or for specific substrate labellings, such as uniform labelling.

The proposed computational methods are efficient and can be used with all network topologies. Thus they can be applied with detailed models of metabolism.

As METAFoR, the methods behind FiatFlux and our method share the idea of augmenting stoichiometric constraints with flux ratios, the methods are complementary and can be applied in tandem. The (constraints to) flux ratios produced by the methods can all be collected together and the fluxes that satisfy all constraints be computed.

Chapter 4

A direct framework for ^{13}C metabolic flux analysis

In this chapter we propose a framework for direct ^{13}C metabolic flux analysis and associate our contributions to the components of this framework. The contributions presented in Part II of thesis will only be described briefly, while new results will be elaborated on more.

4.1 Process of ^{13}C metabolic flux analysis

In the previous chapter approaches for ^{13}C metabolic flux analysis were categorized based on the computational techniques they apply when constructing and solving an equation system constraining the flux distribution. In addition, a complete framework for ^{13}C metabolic flux analysis contains other important steps. In Figure 4.1 the view of a process for direct ^{13}C metabolic flux analysis proposed in this thesis is given (cf. a process view for optimization approach proposed in [WMdG01] and [Wie02]). In the next sections we describe the process in detail.

4.2 Model construction

Before any computational analysis the model of the metabolism of an organism has to be fixed. As described in Chapter 3, in ^{13}C metabolic flux analysis this model consists of the metabolites and the chemical reactions in the metabolic network. For each reaction, carbon mappings from substrates to products and the stoichiometric coefficients describing the molar ratios of substrates and products have to be specified.

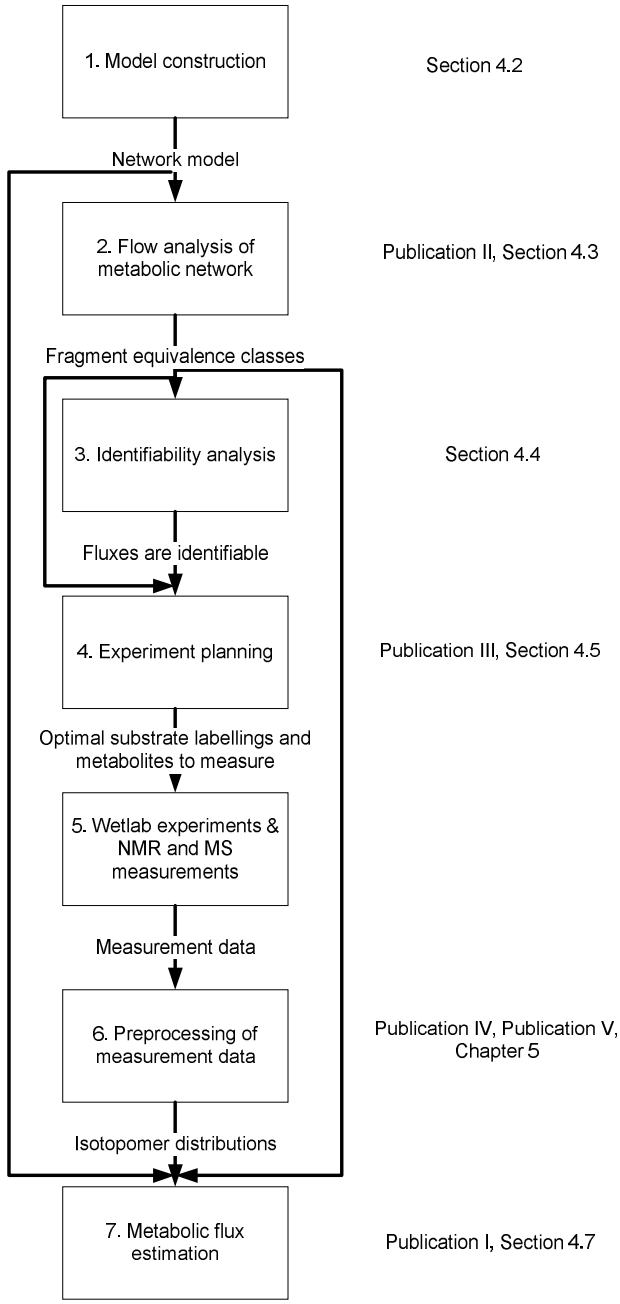


Figure 4.1: Process of ^{13}C metabolic flux analysis.

Public databases such as KEGG LIGAND [GOH+02] contain a lot of biochemical reactions discovered from a multitude of organisms. The ARM database [ARM] augments LIGAND reactions with automatically reconstructed atom mappings [Ari99, Ari03]. It makes sense to utilize this available information to ease the model construction. Unfortunately, matching the reactions of a metabolic network given by a user with reactions in LIGAND and ARM is nontrivial due to diverse naming conventions of metabolites. For example, in a user model *d-xylulose 5-phosphate* might be called "XU5P-D" while in the database a longer name for the same metabolite is used. Furthermore, the user model might contain "glucose" as a carbon source, but in the database reactions might consume "d-glucose", or vice versa.

To solve the problem arising from the different naming conventions we have developed a software tool called ReMatch to assist the construction of metabolic network models. The tool matches the reactions of a user model with the reactions in LIGAND and ARM stored to a local database. In the matching, metabolite names used in the user model are automatically compared with the different synonyms used for metabolites. Synonyms for metabolite names are obtained from LIGAND and from MetaCyc [KZM+04] databases. If the perfect match between a user reaction and a database reaction is not found, a user can easily browse database reactions that partially match the user reaction and select the correct one from the candidates. Furthermore, a user can easily add her own reactions, together with carbon mappings, to the local database. After the matching, ReMatch outputs the stoichiometric matrix of the constructed metabolic network, mappings between the metabolite names used in user model and in the reactions of KEGG and ARM and 13C-FLUX [WMdG01] compatible input file for ^{13}C metabolic flux analysis. The model is also stored to the local database for the methods presented in the thesis. Thanks to ReMatch, it is easy to construct models for ^{13}C metabolic flux analysis as the laborious tasks of metabolite name matching and the construction of the carbon mappings of reactions are handled (mostly) automatically. ReMatch is freely available as a web service at `http://www.cs.helsinki.fi/group/sysfys/software/rematch/index.html`.

4.3 Flow analysis of metabolic network

After its construction, the model of metabolic network is examined with the flow analysis techniques introduced in Publication II. The goal of the flow analysis is to partition the fragments of the metabolites in the network to

equivalence classes such that fragments in the same equivalence class have identical isotopomer distributions in every steady state, when the (joint) isotopomer mappings induced by the network are taken into account (see Figure 4.2 for an example). In abstract terms, equivalence of two fragments follows from their similiar history in the metabolic network. For example, fragments F and F' are necessarily equivalent, if

1. $|F| = |F'|$;

2. all carbons of F originate always from the carbons of F';

3. carbon of F' stay connected to each other via all pathways from F' to F;

4. composite carbon mappings are the same in all pathways from F' to F.

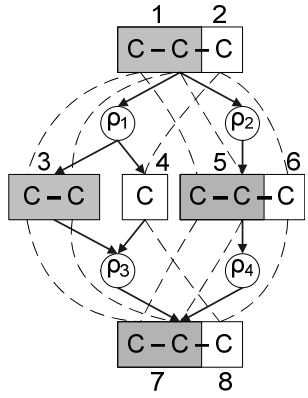


Figure 4.2: An example of equivalence classes of fragments. Grey and white fragments constitute two equivalence classes $\{1, 3, 5, 7\}$ and $\{2, 4, 6, 8\}$. Dashed lines illustrate carbon mappings.

Fragment equivalence classes have many uses (Publication II, Section 5). Most importantly, measured isotopomer constraints to fragment F can be directly propagated to another fragment F' in the same equivalence class, by applying the joint isotopomer mappings between F and F'. This helps in the construction of generalized balance equations (3.7) where isotopomer information is required for each subpool of junction metabolites.

The equivalence classes also facilitate the selection of metabolites that need to be measured to obtain maximal flux information (see Section 4.5) and help in the structural identifiability analysis (Section 4.4) and in the regularization of an equation system constraining the fluxes (Section 4.7.4).

We construct the fragment equivalence classes by first transforming metabolic network G to *fragment flow graph* $\mathcal{F}(G)$ that better models the transfer of metabolite fragments in the network and the cleavage and formation of carbon-carbon bonds during the transfer. $\mathcal{F}(G)$ has connected fragments of all metabolites as nodes. The edges of $\mathcal{F}(G)$ are derived from the reactions of the network: If some reaction maps a substrate fragment $M|F'$ to a product fragment $M'|F$, a directed edge from F' to F is inserted to $\mathcal{F}(G)$. If product fragment $M'|F$ is composed from carbons of more than one substrate molecule by some reaction, a directed edge from root node Δ of $\mathcal{F}(G)$ to F is inserted. Also, there exist directed edges from Δ to all fragments of external substrates in $\mathcal{F}(G)$.

Next the *dominator tree* [AHU74] of $\mathcal{F}(G)$ is constructed. We say that fragment F' dominates, or is a dominator of fragment F in $\mathcal{F}(G)$, if every path from the root node Δ to F goes through F' and the composite carbon mappings defined by all pathways from F' to F in metabolic network G are the same. F' is an immediate dominator of F, denoted by $F' = idom(F)$, if

1. $F \neq F'$,
2. F' dominates F,
3. F' does not dominate any other dominator of F (cf. [App98]).

In dominator tree T corresponding to $\mathcal{F}(G)$ there exists a directed edge from $idom(F)$ to F for each fragment $F \in \mathcal{F}(G)$. As Δ dominates every fragment in $\mathcal{F}(G)$, it is the root of T. Now the fragments belonging to the same subtree of Δ in T are equivalent (Publication II, Theorem 3). Thus, subtrees of T define a partition of fragments in G to the required equivalence classes. Figure 4.3 presents a small metabolic network, the corresponding fragment flow graph and the dominator subtrees corresponding to equivalence classes.

The fragment flow graph $\mathcal{F}(G)$ constructed above is quite large as it contains $O(2^{|M|})$ nodes for each metabolite M. In Section 4 of Publication II we show that it is enough to add only nodes corresponding to one carbon and connected two carbon fragments to $\mathcal{F}(G)$ and construct dominance relations of other fragments by intersecting nodes of dominator tree T corresponding to $\mathcal{F}(G)$ that share a carbon. Thus it is enough to add $O(|M|^2)$ nodes for each metabolite M to $\mathcal{F}(G)$.

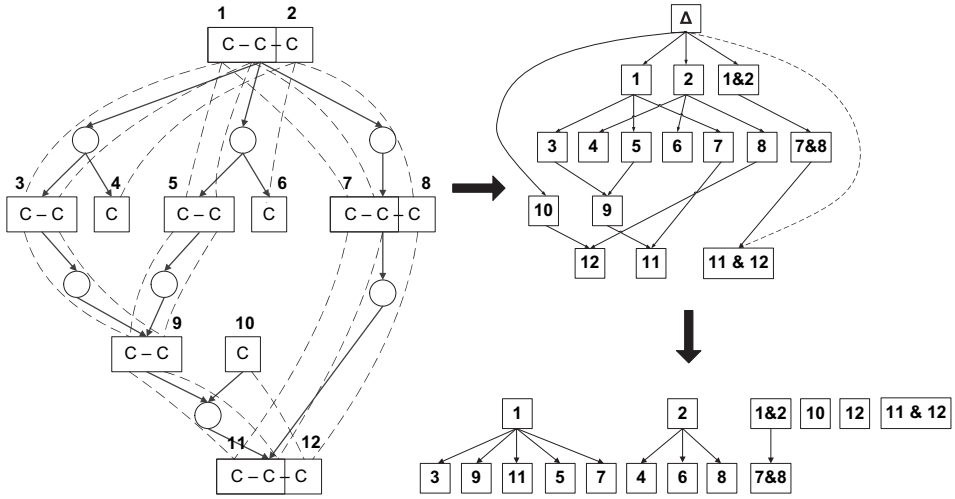


Figure 4.3: A metabolic network (left), the corresponding fragment flow graph (up right) and the subtrees of the dominator tree (down right).

4.3.1 Computation of dominator tree*

Our definition of dominator tree deviates from the general dominator trees [App98] only in the requirement that the composite carbon mappings defined by all pathways from $idom(F)$ to F have to be the same. We call a dominator tree without this additional requirement a weak dominator tree (see Figure 4.4 for illustration). Another way to formulate the requirement of the equal composite carbon mappings is to require that every carbon of F is dominated by a carbon in $idom(F)$. Thus we can apply well-known algorithms [LT79] to first construct a weak dominator tree of fragment flow graph $\mathcal{F}(G)$ and then check for each edge in the weak dominator tree whether it also belongs to the dominator tree by inspecting if the carbons of the dominated fragment are also dominated.

A description of a relatively simple algorithm for the computation of weak dominator tree with the time complexity of $O(m \log n)$, where m is the number nodes and n the number of edges in a flow graph, is given in [App98]. The algorithm is originally due to Lengauer and Tarjan [LT79], who also give an asymptotically faster version with time complexity $O(m\alpha(m, n))$, where α is a (slowly growing) inverse-Ackermann function. Also, more complicated — and in practice slower — algorithms for linear time construction of weak dominator tree [AHWT99, BKRW98] have been presented.

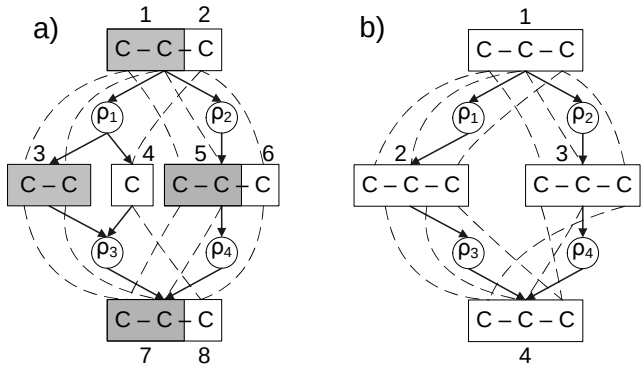


Figure 4.4: Example of weak domination. In a), carbon mappings from fragment 1 to 7 and from fragment 2 to 8 are the same on both pathways. Thus 1 dominates 7 and 2 dominates 8. In b), carbon mappings of reactions ρ_3 and ρ_4 are not the same. Thus metabolite 1 only weakly dominates metabolite 4 (figure from Publication II).

In the Lengaeur-Tarjan algorithm, the concept of semidominator is applied. Let S be a depth-first spanning tree of $\mathcal{F}(G)$ (a recursion tree traversed by the basic depth-first search algorithm). The semidominator of F ($semi(F)$) is a node $F' \in S$ with the smallest depth-first number such that there exists a path from F' to F in $\mathcal{F}(G)$ that does not contain any ancestors of F in S. Now, on the spanning tree path below $semi(F)$ and above or including F, let F' be the node with the smallest numbered semidominator. Then [App98, LT79],

$$idom(F) = \begin{cases} semi(F) \text{ if } semi(F') = semi(F) \\ idom(F') \text{ if } semi(F') \neq semi(F). \end{cases} \tag{4.1}$$

The above conditions can be utilized by visiting the nodes of $\mathcal{F}(G)$ in depth-first order, starting from the node with the highest depth-first number, and gradually constructing a spanning tree and computing semidominators of nodes. The semidominators can be found by noting that a set of candidates for $semi(F)$ consists of 1) predecessors F' of F in $\mathcal{F}(G)$ that are also ancestors of F in S and 2) nodes that are ancestors of F' in S, if F' is a predecessor of F in $\mathcal{F}(G)$, but not its ancestor in S [App98, LT79]. Simultaneously with the computation of semidominators, immediate dominators

that fill the first condition of (4.1) are computed. Finally, the missing immediate dominators are computed based on the second condition of the theorem.

Algorithms 1, 2 and 3 constitute an $O(n \log m)$ time algorithm by Leangauer and Tarjan [LT79] for the computation of a weak dominator tree. The presentation of the algorithm is due to Appel [App98].

4.3.2 Independence analysis of fragments

The fragment flow analysis presented above partitions a metabolic network to equivalence classes of fragments that have identical isotopomer distributions, regardless of the steady state flux distribution of the network. With further analysis it is also possible to know if the isotopomer distribution $D(M_i|E \cup M_i|F)$ of the union of the two fragments $M_i|E$ and $M_i|F$ that do not share carbons depends only from $D(M_i|E)$ and $D(M_i|F)$ but not from the fluxes. This is the case if

1. $M_i|E$ and $M_i|F$ are dominated by some fragments of the metabolic network and

2. the carbons of E have been disconnected from the carbons of F, that is, carbons of E reside in different metabolite than carbons of F, in some stage of all pathways producing M from their dominators.

(This is another example of "similar history" of fragments E and F in metabolic network. See Section 4.3.) If, however, carbons of E and F have stayed intact in some, but not all pathways from dominators of E and F, $D(M|E \cup M|F)$ can depend on the relative fluxes through competing pathways (cf. Figure 2.3 and the end of Section 2.3).

We say that fragment $M|F$ is *intact* in pathway $\mathcal{P}$ if carbons of F are in each step of $\mathcal{P}$ mapped to the same metabolite, i.e. the carbons stay together. If M is not a substrate of $\mathcal{P}$, $M|F$ is intact in $\mathcal{P}$ if and only if a substrate fragment $M'|H$ of $\mathcal{P}$ is intact in pathway $\mathcal{P}' \in \mathcal{P}$ from M' to M, $\mathcal{P}'$ maps H to F, and $M|F$ is intact the rest of the pathway $\mathcal{P}$. In the proof of the following theorem, the concept of fragment is used loosely to aid readability. When we say that carbon e of fragment E resides in a different metabolite than carbons of F in pathway $\mathcal{P}_l$, we mean that e is mapped to E by $\mathcal{P}_l$ and that e belongs to a metabolite in $\mathcal{P}_l$ that does not contain carbons that are mapped to F by $\mathcal{P}_l$.

Theorem 1. *Let $M_i|E$ and $M_i|F$, $E \cap F = \emptyset$, be fragments in the metabolic network G and let $M_p|H$ and $M_q|J$ be some dominators of $M_i|E$ and $M_i|F$, respectively. If there does not exist a pathway $\mathcal{P}_k$ producing $M_i|E \cup F$*

Algorithm 1 An algorithm for computing a weak dominator tree [App98].

Input: flow graph $\mathcal{F}(G)$ and its root Δ
Output: dominator tree T of $\mathcal{F}(G)$

```
function DOMINATORS(F(G),Δ)
    N ← 0
    for n ← 0 to |V(F(G)| − 1 do
        bucket[n] ← {}
        dfnum[n] ← semi[n] ← ancestor[n] ← idom[n] ← null
        samedom[n] ← null
    end for
    DepthFirstNumber(null, Δ)
    for i ← N − 1 downto 1 do
        n ← vertex[i]; p ← parent[n]; s ← p
        for each predecessor v of n do                ▷ compute semidominators
            if dfnum[v] ≤ dfnum[n] then
                s' ← v
            else
                s' ← semi[ancestorWithLowestSemi]
            end if
            if dfnum[s'] < dfnum[s] then
                s ← s'
            end if
        end for
        semi[n] ← s
        bucket[s] ← bucket[s] ∪ n
        Link(p, n)
        for each v ∈ bucket[p] do
            y ← AncestorWithLowestSemi(v)
            if semi[y] = semi[v] then                 ▷ first clause of 4.1
                idom[v] ← p
            else
                samedom[v] ← y
            end if
        end for
        bucket[p] ← {}
    end for
    for i ← 1 to N − 1 do
        n ← vertex[i]
        if samedom[n] ≠ null then                     ▷ second clause of 4.1
            idom[n] ← idom[samedom[n]]
        end if
    end for
    return idom
end function
```

Algorithm 2 An algorithm for computing depth-first numbering to nodes of $\mathcal{F}(G)$.

Input: root Δ of flow graph $\mathcal{F}(G)$
Output: depth-first numbering of $\mathcal{F}(G)$

```
function DepthFirstNumber(p, n)
    if dfnum[n] = 0 then
        dfnum[n] ← N; vertex[N] ← n; parent[n] ← p
        N ← N + 1
        for each successor w of n do
            DepthFirstNumber(n, w)
        end for
    end if
end function
```

Algorithm 3 Algorithms for finding an ancestor of node v in a spanning tree, represented by an array *ancestor*, that has the semidominator with lowest depth-first number and for maintaining the spanning tree.

```
function AncestorWithLowestSemi(v)
    a ← ancestor[v]
    if ancestor[a] ≠ null then
        b ← AncestorWithLowestSemi(a)
        ancestor[v] ← ancestor[a]
        if dfnum[semi[b]] < dfnum[semi[best[v]]] then
            best[v] ← b
        end if
    end if
    return best[v]
end function
function Link(p, n)
    ancestor[n] ← p
    best[n] ← n
end function
```

from the external substrates such that carbon pair (a, b)*, where* $a \in E$ *and* $b \in F$*, stays intact in a subpathway of* $\mathcal{P}_k$ *from* M_p *or* M_q *to* M_i*, then* $D(E \cup F) = D(E) \otimes D(F)$*.*

Proof. By the definition of domination, all pathways producing M_i from the external substrates of G produce $M_i|E$ from $M_p|H$ and $M_i|F$ from $M_q|J$. Furthermore, in all pathways $\mathcal{P}_l$ from the external substrates to M_i, H (resp. J) is transported intact to E (F) and the joint carbon mappings between H and E (J and F) are the same. We assume that in some stage including and after M_p or M_q of all pathways $\mathcal{P}_l$, every carbon of E resides in the different metabolite than all carbons of F. In M_i, however, E and F reside in the same metabolite. Thus, for every $\mathcal{P}_l$, there exists a reaction ρ_l that combines E and F to fragment $E \cup F$ of metabolite M_y such that $E \cup F$ travels intact to M_i. Thus, $D(E \cup F) = D(E) \otimes D(F)$ for M_i molecules produced by $\mathcal{P}_l$ (Assumption 5 in Section 3.1). Because $D(E \cup F) = D(E) \otimes D(F)$ holds for molecular fragments $E \cup F$ produced by every pathway $\mathcal{P}_l$, it holds also for the total pool of M_i. □

Theorem 1 gives us tools for more efficient propagation of isotopomer data: it allows us to derive new constraints to fragment $M_i|E \cup M_i|F$, if some constraints to $D(E)$ and $D(F)$ are known. Let $M_p|H$ and $M_q|J$ be dominators of $M_i|E$ and $M_i|F$ that dominate all other dominators $M_i|E$ and $M_i|F$. The applicability of Theorem 1 can be checked with a relatively simple algorithm: First we construct a graph G^p that has a labelled node corresponding each carbon pair (a, b) of every metabolite and its inflow subpools. If there exists a reaction in the metabolic network where carbons a and b of metabolite M_j are mapped to carbons a' and b' of metabolite M_k a directed edge $(a, b) \rightarrow (a', b')$ is inserted to $\mathcal{G}$. Now, if there exists a path in $\mathcal{G}$ from M_p or M_q to node (a, b) in G^p, the label $L(a, b)$ equals "connected". Otherwise $L(a, b)$ equals "disconnected". Then, $D(F \cup F') = D(F) \otimes D(F')$ if there does not exist a carbon pair $(a, b), a \in F \wedge b \in F'$, whose state equals "connected".

4.4 Structural identifiability analysis

The (structural) identifiability analysis of a metabolic network tries to find out how much information about the fluxes can be obtained from the available measurements, in the best case. The goal of the analysis is to study whether the measurements have potential to reveal the fluxes of interest. If not, it is not worthwhile to conduct wet lab experiments. Identifiability analysis is related to the experiment planning of the next section: one

should plan such experiments so that maximal flux information is potentially obtainable with reasonable measurement cost.

For optimization approach of ^{13}C metabolic flux analysis, techniques for identifiability analysis are given in [vWHVG01] and [IW03]. In the context of direct ^{13}C metabolic flux analysis also identifiability analysis can be tackled with the help of equivalence classes. The flux distribution can be fully solved with direct methods only if the linear equation system $A\mathbf{v} = e$ containing mass balances (2.1) and generalized isotopomer balances (3.7) contain as many linearly independent constraints as there are fluxes in the model. In other words, coefficient matrix A has to be of full rank. If we have n fluxes and the rank of the stoichiometric matrix containing (2.1)'s equals k, we need at least $(n-k)$ more constraints from isotopomer measurements to obtain $rank(A) = n$.

In the next subsection we derive upper bounds to the number of independent flux constraints (3.7) written for each junction M_i by analyzing the equivalence classes. If the total sum of these constraints in the network is less than $(n-k)$, we know that all fluxes cannot be solved with the help of isotopomer data alone (cf. [vWHVG01]). The same upper bounds for independent flux constraints are also useful in the stability analysis of an equation system constraining the fluxes (see Section 4.7.4).

4.4.1 Upper bounds to flux information of generalized isotopomer balance equations*

Let M_i be a junction metabolite produced by k reactions.

Observation 1 ([vWHVG01])**.** *Isotopomer balances (3.7) for M_i contain at most $(k-1)$ constraints that are independent from the mass balance (2.1) of M_i. These constraints can give new information only about the relative fluxes of reactions producing M_i, not about the relative fluxes of reactions consuming M_i.*

Proof. From the construction of balance equations (2.1) and (3.7) we see that in the coefficient matrix A_i of the linear equation system containing the balance equations (2.1) and (3.7) for M_i, columns corresponding to the reactions consuming M_i are linearly dependent from the columns corresponding reactions producing M_i. Thus, balances (2.1) and (3.7) can together contain at most k linearly independent constraints to M_i and, compared to the mass balance (2.1), isotopomer balances (3.7) can give new independent constraints only to the fluxes of the reactions producing M_i. □

Observation 2 ([vWHVG01])**.** *Isotopomer balances (3.7) for M_i contain at most $2^{|M_i|} - 1$ constraints that are independent from the mass balance (2.1) of M_i.*

Proof. There exists $2^{|M_i|}$ components in the isotopomer distribution of M_i. As the sum of abundances of all isotopomers is equal to 1, the mass balance (2.1) and $2^{|M_i|} - 1$ linearly independent isotopomer balances fix the value of the remaining isotopomer balance. □

Observation (1) can be generalized by taking into account the partition of metabolite fragments to equivalence classes. Let $\mathcal{P}_F$ be a partition of subpools $M_{ij}|F$ of fragment $M_i|F$ to subsets such that $M_{ij}|F$ and $M_{ik}|F$ belong to the same subset if and only if $M_{ij}|F$ and $M_{ik}|F$ are equivalent. Let $|\mathcal{P}_F|$ denote the number of subsets in $\mathcal{P}_F$.

Observation 3. *Isotopomer balances (3.7) based on isotopomer constraints known for subpools of $F' \subseteq F$ contain at most $|\mathcal{P}_F| - 1$ constraints to the fluxes producing M_i that are independent from the mass balance (2.1).*

Proof. If $M_{ij}|F$ and $M_{ik}|F$ belong to the same equivalence class, coefficients d_{ijh} and d_{ikh} of (3.7) are always identical. Thus, in the coefficient matrix A_i of the linear system containing the balance equations (2.1) and (3.7) for M_i, the column corresponding to reaction ρ_j is linearly dependent from the column corresponding to reaction ρ_k. From this and Observation 1 the result follows. □

Corollary 1. *If (mixed pool) $M_i|F$ is dominated by some fragment, isotopomer balances (3.7) of $M_i|F$ reduce to mass balance (2.1) of M_i.*

Observation 2 can also be generalized by taking the equivalence classes into account. If $M|F$ has a dominator fragment and no $E \supseteq F$ has a dominator fragment, we say that F is a *maximal dominated fragment*. Let $\mathcal{D}_i$ contain all maximal dominated fragments of M_i (here also maximal fragments obtained from the independence analysis (Section 4.3.2) are considered as maximal dominated fragments but their subfragments are not). For each fragment $M_i|F_l \in \mathcal{D}_i$, let C_l denote a binary matrix of size $2^{|F_l|} \times 2^{|M_i|}$. Each row of C_l corresponds to a component in the isotopomer distribution of F_l and column to a component in the isotopomer distribution of M_i. $C_l(m, n) = 1$ if and only if $M_i(n)$ belongs to $F_l(m)$. Let matrix C be constructed by vertically concatenating matrices C_l . Let $\mathcal{C}$ denote a subspace of $\mathcal{I}_{Mi}$ spanned by C.

Theorem 2. *Isotopomer balances (3.7) for M_i contain at most*

$$2^{|M_i|} - \sum_{F_l \in \mathcal{D}_i} 2^{|F_l|} - 1$$

constraints to the fluxes producing M_i that are independent from the mass balance (2.1) of M_i.

Proof. From Corollary 1 we know that the projection of the isotopomer constraints of subpools M_{ij} to subspace $\mathcal{C}_l$, spanned by any C_l, does not contain any flux information additional to the mass balance (2.1) of M_i. Then, also the projection of the isotopomer constraints of subpools M_{ij} to subspace $\mathcal{C}$ can not contain any flux information additional to (2.1). This is because the projection is a linear combination of isotopomer constraints projected to subspaces $\mathcal{C}_l$ (maximal dominated fragments F_l do not share carbons). Thus, only the projection of isotopomer information of M_{ij}'s to the orthogonal complement $\mathcal{C}\perp$ of $\mathcal{C}$ can contain flux information. The rank of $\mathcal{C}\perp$ equals $2^{|M|} - rank(C)$.

As fragments F_l do not share carbons, rows of matrices C_l are all linearly independent. Thus $rank(C) = \sum_{F_l \in \mathcal{D}_i} 2^{|F_l|}$. □

4.5 Planning carbon labelling experiments

As stated in Section 3.2 it is very nontrivial to measure isotopomer distributions of intermediate metabolites with current technologies. In (tandem) mass spectrometer, it is necessary to separate metabolites in the cell extract. Also, for each metabolite or metabolite group a specific experimental protocol has to be developed. In NMR, spectral overlap, large differences in concentration and available spectral edition techniques make the labelling information of some metabolites more accessible than others. Furthermore, the labelling of external substrates has a great effect on the flux information that is obtained from the isotopomer measurements.

Thus it is worthwhile to carefully plan the isotopomer tracer experiments before conducting them. The planning has two general questions to answer: 1) What is the optimal labelling of external substrates for flux estimation? 2) Which metabolites of the network are the most important to measure, that is, which metabolites carry the most flux information in their isotopomer distribution. Question 1) is studied in the context of optimization approach for ^{13}C metabolic flux analysis in [MWKdG99] and in [ABS03].

Answers to the question 2) help us to select small subsets of metabolites to measure that give us enough information to solve the fluxes of the

network and thus reduce the experimental effort needed. In [AKS06], this question is touched in the context of optimization approach by studying which measurements contribute significantly to the variance of the estimated fluxes. Publication III concentrates on this question in the context of a direct approach for ^{13}C metabolic flux analysis. In Publication III we model the situation at the beginning of the process of the development of measurement technologies. In that stage we do not yet know, what kind of constraints are in practice measurable for each metabolite in the network but want to concentrate the development of measurement technologies on the most promising metabolites. In this situation we need to make some assumptions about the quality of the measurement data eventually available. In Publication III we assume that either positional enrichments, that is, the labelling degrees of (some) carbons of (some) metabolites in the network or the full isotopomer distributions are in principle measurable. With these assumptions we can model the problem of selecting a minimal set of informational metabolites to measure as a set cover problem [ACK$^+$99]. As shown in Section (4.7.1), isotopomer constraints can be propagated from one fragment to another inside the same equivalence class. Thus to obtain as many (3.7)'s as the measurement of every metabolite in the network would allow, every carbon c_k in each junction metabolite and in their inflow subpools has to be "covered" with at least one measurement from the equivalence class of c_k. In other words, at least one fragment from each equivalence class containing junction fragments has to be measured.

4.6 Cultivations, measurements and preprocessing of measurement data

If the identifiability analysis shows that it is possible to obtain the required flux information, cultivations are carried out with the planned labelling of input substrates. Then samples are collected and prepared for NMR and mass spectrometry measurements. NMR and mass spectrometry do not directly output constraints to the isotopomer distribution of the carbon chain of a measured metabolite, as required by (3.7). Thus the raw measurement data has to be preprocessed before it can be utilized in ^{13}C metabolic flux analysis. For NMR, computational tools for obtaining isotopomer constraints from raw spectra are introduced in [WMWdG96],[SGH$^+$99] and [vWSVH01]. Preprocessing of MS-MS data for ^{13}C metabolic flux analysis is discussed more thoroughly in the next chapter.

4.7 Metabolic flux estimation

After the raw measurement data is converted to linear constraints to the isotopomer distributions of metabolites in the network, we are finally ready to estimate the flux distribution of a metabolic network. In our direct method for ^{13}C metabolic flux analysis the flux estimation algorithm divides into three main steps:

1. propagation of the measurement data,
2. construction of the linear equation system constraining the fluxes and
3. solving the equation system and analyzing the sensitivity of the result.

4.7.1 Propagation of measurement data

The aim of the propagation of measurement data is to infer from the isotopomer constraints of measured metabolites as many isotopomer constraints as possible to unmeasured metabolites. As a rule of thumb, more constraints the unmeasured metabolites will get, more generalized isotopomer balance equations (3.7) bounding the fluxes can be written. By definition, fragment equivalence classes can be utilized in the measurement propagation. From isotopomer constraints known for fragment $M_i|F$ isotopomer constraints for other fragments $M_l|F_k$ in the equivalence class of F can be computed by applying joint isotopomer mappings defined by pathways between F and F_k.The equivalence classes can contain fragments from the opposite sides of a junction metabolite and thus facilitate the propagation of isotopomer information also through junction metabolites (see Figure 4.3 and Section 5 of Publication II). Thus the measurement propagation with the help of equivalence classes improves the basic propagation method given in Publication I.

Before measurements can be propagated from fragment $M|F$ of measured metabolite M to other fragments in the equivalence class of F, we need to infer isotopomer constraints to F from the constraints measured to the whole metabolite M. In this projection of measurements from M to $M|F$ we want to avoid any unnecessary loss of measurement information, that is, we want to obtain as many linearly independent constraints to the isotopomer distribution of F as possible. For example, if we have measured that a two-carbon metabolite M has the isotopomer distribution $P_M(00) = 0.2, P_M(01) = 0.3, P_M(10) = 0.4, P_M(11) = 0.1$ and require isotopomer constraints for fragment $M|F$ consisting of the first carbon of M, we should obtain $P_F(0) = P_M(0*) = P_M(00) + P_M(01) = 0.5$ and $P_F(1) = P_M(1*) = P_M(10) + P_M(11) = 0.5$.

For the general model of isotopomer measurements (2.8) the projection of measurement information from a metabolite to its fragments can be done by the techniques of linear algebra introduced in Publication I. These techniques are recapitulated in the next subsection.

Projecting measurements to fragments

Let S_i denote a matrix with $2^{|M_i|}$ columns, one column for each isotopomer b of M_i. Each row h of S_i corresponds to a measured isotopomer constraint (2.8) such that $S_i(h, b) = s_{b,i,h}$. Now the rows of S_i span subspace $\mathcal{S}_i \subseteq \mathcal{I}_{Mi}$ where the measurement data for M_i lie (See Section 3 of Publication I). On the other hand, let U_k denote a matrix with also a column for each isotopomer $M_i(b)$ and a row for each isotopomer $F_k(b')$ of $M_i|F_k$, that is,

$$U_k(b', b) = \begin{cases} 1 & \texttt{if } b_j = b'_j \texttt{ for all carbon positions } j \in F_k \\ 0 & \texttt{otherwise.} \end{cases} \tag{4.2}$$

Now the rows of U_k span another subspace $\mathcal{U}_k \subseteq \mathcal{I}_{Mi}$ where isotopomer constraints for F_k have to lie. To obtain isotopomer constraints for fragment $M_i|F_k$ from a measurement $S_i D(M_i) = \mathbf{d}_i$, we need to compute the vector space intersection $\mathcal{Y}_{i,k} = \mathcal{S}_i \cap \mathcal{U}_k$ and project the measurement to $\mathcal{Y}_{i,k}$. This can be done by applying the techniques of linear algebra described in more detail in Section 3.1 of Publication I and Section 3.2, A.1 and A.2 of [RRU02].

Let $Y_{i,k}$ be a matrix with row space $\mathcal{Y}_{i,k}$. As soon as we know isotopomer constraints

$$Y_{i,k} D(M_i|F_k) = \mathbf{d}_{f_k}$$

to fragment F_k of a measured metabolite M_i, we can easily propagate the projected measurement to other fragments $M_j|F_l$ in the equivalence class of F_k: we only need to compute a (composite) isotopomer mapping matrix $IMM^{k,l}$ from F_k to F_l by successively applying isotopomer mapping matrices in some pathway between F_k and F_l and then multiplying $Y_{i,k}$ by $IMM^{k,l}$ to obtain $Y_{j,l}$ such that $Y_{j,l} D(M_j|F_l) = \mathbf{d}_{f_k}$.

After the propagation of measurement data inside the fragment equivalence classes, new isotopomer constraint for unions of some fragments of the same metabolite can be derived, as described in Section 4.3.2.

4.7.2 Construction of generalized isotopomer balances

In the second step of our metabolic flux estimation algorithm a linear equation system containing flux constraints obtained from mass balances (2.1)

and generalized isotopomer balances (3.7) is constructed. After the propagation step of Section 4.7.1 we have some isotopomer constraints $S_{ij}D(M_{ij}) = \mathbf{d}_{ij}$ for each subpool j of every junction metabolite M_i. (For non-junction metabolites, isotopomer balance equations do not contain any additional flux information compared to the mass balances. See Observation 1.) In the best case we know complete isotopomer distribution $D(M_{ij})$, in the worst case we have only trivial constraints stating that the sum of relative abundances of all isotopomers equals one. However, the isotopomer constraints of different subpools do not yet conform to (3.7) as the matrices S_{ij} are not necessarily the same. Thus we still need to compute a common subspace $\mathcal{Y}_i = \bigcap_j \mathcal{S}_{ij}$ ($\mathcal{S}_{ij}$ is spanned by the rows of S_{ij}) of the isotopomer constraints known for each subpool M_{ij} and project subpool constraints $S_{ij}D(M_{ij}) = \mathbf{d}_{ij}$ to $\mathcal{Y}_i$ (see Section 3.3 of Publication I). This can be done with the same techniques that were applied to project measured isotopomer information of a metabolite to its fragments in Section 4.7.1. Let Y_i be a matrix with row space $\mathcal{Y}_i$. After the projection we obtain isotopomer constraints $Y_iD(M_{ij}) = \mathbf{z}_{ij}$ for each subpool M_{ij}. Now the isotopomer constraints of all the subpools lie in the same subspace of $\mathcal{I}_{Mi}$ and we are ready to write the system of generalized isotopomer balance equations (3.7) for every junction M_i :

$$\sum_{j=1}^{n} \alpha_{ij} v_j \mathbf{z}_{ij} = \beta_i \mathbf{z}_{i0}, \tag{4.3}$$

that is (cf. Equation (8) of Publication I),

$$A_iV = \begin{bmatrix} \alpha_{1i}(\mathbf{z}_{1i})_1 & \cdots & \alpha_{ni}(\mathbf{z}_{ni})_1 \\ \vdots & \ddots & \vdots \\ \alpha_{1i}(\mathbf{z}_{1i})_r & \cdots & \alpha_{ni}(\mathbf{z}_{ni})_r \end{bmatrix} \cdot \begin{bmatrix} v_1 \\ \vdots \\ v_n \end{bmatrix} = \mathbf{g}_i, \tag{4.4}$$

where $\mathbf{g}_i = \beta_i \mathbf{z}_{i0}$. (As a trivial isotopomer constraint stating that isotopomer abundances sum to 1 is contained in generalized balance equation, we do not need to explicitly add mass balance (2.1) to (4.4) of a junction.)

When (4.4)'s of all junction metabolites are combined with the mass balances (2.1) of non-junctions, we obtain a linear equation system

$$A\mathbf{v} = \begin{bmatrix} A_1 \\ \vdots \\ A_m \end{bmatrix} \cdot \begin{bmatrix} v_1 \\ \vdots \\ v_n \end{bmatrix} = \begin{bmatrix} \mathbf{g}_1 \\ \vdots \\ \mathbf{g}_m \end{bmatrix} = \mathbf{g} \tag{4.5}$$

constraining the fluxes $\mathbf{v}$ of the network that contains a block (junctions) or a row (non-junctions) A_k for each metabolite M_k.

4.7.3 Solving the system

If (4.5) is of full rank, the fluxes can be (in principle) solved with standard linear algebra [Sch89] (See also 4.7.4). If the system is of less than full rank, a single flux distribution can not be pinpointed. Instead, (4.5) defines the space of feasible flux distributions, that are all equally good solutions. In that case we can apply techniques developed for the analysis of stoichiometric matrices to determine as many fluxes as possible [KS02] from (4.5). More generally, by linear programming we can obtain maximum (resp. minimum) values for each flux v_i:

$$\begin{aligned} &\texttt{For each } v_i: \\ &\qquad \max \quad v_i \\ &\qquad \text{s.t.} \quad A\mathbf{v} = \mathbf{g} \\ &\qquad v_i^{min} \leq v_i \leq v_i^{max} \ \forall v_i \in \mathbf{v}, \end{aligned} \tag{4.6}$$

where v_i^{min} and v_i^{max} are predetermined minimum and maximum allowable values for v_i By altering v_i^{min}'s and v_i^{max}'s it is possible to see how different hypotheses about the value of an unconstrained flux v_i affect to the feasible values of the other fluxes.

Furthermore, it is possible to search for an optimal flux distribution from the feasible space defined by (4.5) by solving a linear program analogous to (2.3). In that case isotopomer data constrain the feasible space more than the stoichiometric information would alone do, thus possibly allowing more accurate estimations of the real flux distribution.

4.7.4 Stability analysis

When equation systems based on the measurement data are solved, some estimates on the effect of measurement errors to the result are required. In this subsection we first show how to decrease the effect of measurement errors by regularization techniques based on the singular values and condition numbers of the coefficient matrices defining the systems. Then we sketch a conceptually simple Monte Carlo method to assess the sensitivity of estimated fluxes to measurement errors.

Regularization

In (4.5) coefficients of A originate from isotopomer measurements and components of $\mathbf{g}$ from measured external flows of metabolites. Thus both A and $\mathbf{g}$ inevitably contain measurement errors. The errors in isotopomer measurements may make linearly dependent constraints only "almost linearly

dependent" and transform an underdetermined A to fully (or less under-) determined, nearly singular matrix.

Unfortunately, if a linear system is nearly singular, small perturbations on data can affect greatly to the solution of the system. Furthermore, due to measurement errors the solution produced by numerical computations can be far from the correct one if the system is solved with standard methods [Mol04]. Thus we need to regularize (4.5) before solving it.

The need for regularization can be illustrated with a small example consisting of four reactions and metabolites shown in Figure 4.5. Let $D(M_1) =$

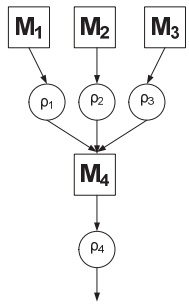


Figure 4.5: An example network.

$D(M_{41}) = [0.3, 0.2, 0.4, 0.1]$, $D(M_2) = D(M_{42}) = [0.2, 0.1, 0.5, 0.2]$ $D(M_3) = D(M_{43}) = [0.1, 0.5, 0.1, 0.3]$ and let the flux $v_1 = v_2 = 1$, $v_3 = 2$ and $v_4 = 4$. Now $D(M_4) = [0.175, 0.325, 0.275, 0.225]$. Let us assume that we are able to measure the mass isotopomer distributions of metabolites, as well as flux v_4. The isotopomer measurements should give $D(M_1)^m = [0.300, 0.600, 0.100]$, $D(M_2)^m = [0.200, 0.600, 0.200]$, $D(M_3)^m = [0.100, 0.600, 0.300]$ and $D(M_4)^m = [0.175, 0.600, 0.225]$. Now (4.4) can be written as

$$A^*\mathbf{v} = \begin{bmatrix} 0.300 & 0.200 & 0.100 & -0.175 \\ 0.600 & 0.600 & 0.600 & -0.600 \\ 0.100 & 0.200 & 0.300 & -0.225 \\ 1 & 1 & 1 & -1 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \\ v_3 \\ v_4 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 4 \end{bmatrix}, \qquad (4.7)$$

where the first three constraints originate from isotopomer measurements and the fourth states the mass balance (2.1). The rank of A equals 3, thus (4.7) is underdetermined. Let us assume that the measurements introduce

small errors to mass isotopomer distributions. So instead of (4.7) we obtain

$$A\mathbf{v} = \begin{bmatrix} 0.310 & 0.190 & 0.09 & -0.177 \\ 0.580 & 0.610 & 0.605 & -0.595 \\ 0.110 & 0.200 & 0.295 & -0.228 \\ 1 & 1 & 1 & -1 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \\ v_3 \\ v_4 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 4 \end{bmatrix}, \tag{4.8}$$

Now the rank of A equals 4 and (4.8) seems to be fully determined (and inconsistent). When (4.8) is solved numerically in Matlab 6.5 environment, a result $v_1 = 1.8$, $v_2 = -0.67$, $v_3 = 2.82$, $v_4 = 3.99$ is obtained. The residual of the given result is small, but still the result is not the correct flux distribution.

There exists many methods regularize a linear equation system [Neu98]. One of the conceptually simplest of these methods is singular value truncation. In short, by analyzing the singular values of coefficient matrices A_i of (4.4) it is possible to detect the situations where the rank of A_i is higher than it should be due to almost dependent constraints caused by measurement noise. These almost dependent constraints can then be removed from (4.4). The removal decreases the rank of A_i but makes (4.4) more tolerant to measurement noise. In the next subsection the technique of singular value truncation is described in more detail.

Singular value truncation*

The singular value decomposition of matrix $A \in \mathbb{R}^{m,n}$, $m \geq n$, is of the form

$$A = U\Sigma W^T = \sum_{i=1}^{n} \mathbf{u}_i \sigma_i \mathbf{w}_i^T, \tag{4.9}$$

where U and W are orthonormal matrices, $UU^T = WW^T = I_n$ and where Σ is a diagonal matrix with diagonal elements $(\sigma_1, \ldots, \sigma_n)$ such that $\sigma_1 \geq \sigma_2 \geq \ldots \sigma_n \geq 0$. The numbers σ_i are the singular values of A. The columns $\mathbf{u}_i | \sigma_i > 0$ of U span the range of A while the columns $\mathbf{w}_i | \sigma_i = 0$ span the null space of A [PTV92]. Thus the rank of A is equal to the number of non-zero singular values of A. The condition number κ of A is defined as

$$\kappa(A) = \sigma_1 / \sigma_n. \tag{4.10}$$

Large but finite $\kappa(A)$ indicates that A is ill-conditioned: there exists columns of A that are nearly linearly dependent and $Ax = b$ is essentially underdetermined, but due to measurement errors etc. it looks like fully

determined. When the equation system is solved without taking care of ill-conditioness, the nearly dependent constraints pull the solution towards the infinity in the direction almost identical to some vector from the null space of A [PTV92].

The idea of singular value truncation is to transform an equation system $Ax = b$ containing an ill-conditioned matrix A to another system $A'x = b$ such that the new system does not contain nearly dependent constraints. The goal is that the rank of A' represent the "true rank" of A, when measurement errors are discarded. The closest approximation A'_k of A, $rank(A') = k$ in least squares sense can be obtained by taking the singular value decomposition $A = U\Sigma W^T$, zeroing singular values $\sigma_{k+1}, \ldots, \sigma_n$ of Σ and then multiplying back $A'_k = U\Sigma W^T$. Furthermore, least squares estimate x^* for $A'x = b'$ can be directly computed from singular value decomposition by zeroing the small singular values and noting that

$$x^* = W[diag(1/\sigma_i)]U^T b, \tag{4.11}$$

where $[diag(1/\sigma_i)]$ denotes a diagonal matrix with diagonal values $1/\sigma_i$. If $\sigma_i = 0$, the corresponding diagonal element is set to zero.

Singular value truncation by zeroing the small singular values can also be used to regularize matrices with fewer constraints than unknowns. However, an open question remains: how to decide what is a cut-off for "too small" singular values. Unfortunately, no simple answer to this instance of noise versus signal question exists. As a general rule of thumb we can try to find a large gap between successive singular values and decide that the gap defines the threshold. For example, the singular values of (4.8) are $[2.46, 0.85, 0.20, 0.01]$. The last singular value σ_4 is considerably smaller than the others, suggesting that the corresponding vector $\mathbf{u_4} \in U$ represents noise and σ_4 should be zeroed. There also exists more general methods, such as the computation of L-curves [HO93], to find a good cut-off threshold.

Interestingly, Observations 1, 2 and 3 and Theorem 2 immediately give more domain specific upper bounds to the rank of (4.4) and thus act as safe cut-off thresholds. For example, in (4.8) junction metabolite M_4 was produced by three reactions. Thus according to Observation (1), $rank(A) \leq 3$ in (4.8).

The singular value truncation technique introduced above to regularize (4.4) constraining the fluxes of a junction can also be applied to regularize (4.5) that constraints the complete flux distribution.

Sensitivity analysis

For experimentalist, it is very important to know how sensitive the obtained estimation of fluxes is to measurement errors. In the previous section we showed how to decrease the sensitivity by regularization, in this section we give techniques for sensitivity analysis that can be applied after regularization. In the non-linear optimization framework for ^{13}C metabolic flux analysis, mathematically involved methods to obtain local linearized estimations for the covariance matrix, standard deviations and confidence intervals of the estimated fluxes [ABS03, DBS01, WSdGM97] or nonlinear heuristics [AKS06] to obtain more accurate estimates on the confidence intervals have been developed. As our direct method for ^{13}C metabolic flux analysis is computationally relatively efficient, we can afford to a simple, yet powerful Monte Carlo procedure to obtain estimates on the variability of individual fluxes:

1. For each measured metabolite M_i: By studying the variability in the repeated measurements, fix the distribution Ω_i from which the measurements of M_i are sampled.

2. Repeat k times:

   (a) For each measured metabolite M_i: sample a measurement from Ω_i.

   (b) Estimate fluxes $\mathbf{v}_l$ from the sampled measurements.

3. Compute appropriate statistics from the set $V = \{\mathbf{v}_1, \ldots, \mathbf{v}_k\}$ to describe the sensitivity of fluxes to measurement errors.

Possible statistics that can be applied in the last step of the above algorithm include standard deviation, empirical confidence intervals [AKS06], kurtosis, standard error etc. of each individual flux v_j and measures of "compactness" of V, such as (normalized) average distance of items of V from the sample average.

4.8 Experiments

In this section we present the initial results of a proof-of-concept demonstration where the direct method of ^{13}C metabolic flux analysis introduced above was applied to estimate the fluxes of the central metabolism of *Saccharomyces cerevisiae*. The goal of this demonstration is to show the practical feasibility of the method.

compound	# of observed carbons
alanine	2
arginine	1
aspartate	2
glutamate	3
glycine	1
histidine	3
isoleucine	5
leucine	5
lysine	5
methionine	1
phenylalanine	2
proline	4
serine	2
threonine	3
tyrosine	3
valine	4

Table 4.1: Detected amino acids and the number of observed carbons in 2D[^{13}C ,^{1}H] COSY spectrum for each measured amino acid .

Isotopomer data for the demonstration originated from an experiment where *S. cerevisiae* was grown in a glucose-limited continuous cultivation on minimal medium [VSvD92]. After reaching the steady state, controlled by constant physiological parameters, 10% of the carbon source glucose was replaced by fully labelled glucose for approximately 1.5 residence times, so that about 78% of the biomass was labelled. 2D[^{13}C ,^{1}H] COSY spectra of the hydrolyzed biomass sample were acquired with Varian Inova 600 MHz NMR spectrometer. The software FCAL v.2.3.0 [SGH^{+}99] was used to compute constraints (2.5) to the isotopomer distributions of 16 amino acids from the spectra. Table 4.1 lists the detected amino acids. For each amino acid, a number of carbons observed with NMR is also listed (cf. Section 2.4). In the computational analysis we used a slightly modified model of the central carbon metabolism of *S. cerevisiae* from [BKS05] (See Figure 4.6). Simplified pathways producing amino acids were taken from [MFC^{+}01]. The model was constructed with ReMatch (Section 4.2), the carbon mappings of the reactions were provided by the ARM project [ARM]. Bidirectional reactions were modelled as two separate reactions. For technical reasons, each amino acid was represented by two metabolite nodes, inter-

nal and external. Cofactor metabolites that do not exchange carbons with primary metabolites and carbon dioxide were excluded from the model. The relative abundances of proteinogenic amino acids and the precursor requirements for the yeast biomass synthesis were obtained from the literature [GMdSCN01, LH01, Our72]. This data was used as an additional constraint to the stoichiometry. Furthermore, the measured consumption of glucose and the production of ethanol were used to further constrain the stoichiometry. After the addition of these constraints, the rank of the stoichiometric matrix of the model was 86. In total, the model contained 104 reactions and 86 metabolites in three cellular compartments (cytosol, mitochondria, external). Thus, there remained $(104 - 86 = 18)$ degrees of freedom in the equation system defined by the augmented stoichiometric matrix.

In the fragment flow analysis of the metabolic network, an extra "labelling node" modelling the equal labelling degree of every carbon of external substrate glucose was inserted to the fragment flow graph between start node Δ and every carbon of glucose, as sketched in Section 5 of Publication II. The fragment flow analysis revealed four completely dominated junction metabolites. From these completely dominated junction metabolites, acetaldehyde and acetate reside in the same unbranched pathway from pyruvate to acetyl-CoA, in cytosol. Acetaldehyde and acetate are junctions only because of the bidirectionality of the reactions that consume the metabolites in the pathway. For these reactions, it is impossible to estimate separate forward and backward fluxes. We removed these backward reactions from the model and treated the fluxes of remaining forward reactions as net fluxes through the metabolites. Thus, there remained 102 fluxes in the model. When constructing the equation system (4.5), the coefficient matrices of equation systems (4.4) constraining the fluxes of the same junction were regularized by applying upper bounds to flux information (Section 4.4.1) as described in Section 4.7.4. Furthermore, the singular values of the coefficient matrices of (4.4)'s, (4.5) and the matrices representing the basis of a common isotopomer subspace $\mathcal{Y}_i$ known for the isotopomer constraints of the each subpool of M_i (Section 4.7.2) that were smaller than a predetermined cut-off threshold were truncated. Figure 4.7 depicts the rank of (4.5) with different cut-off thresholds for small singular values.

We also compared the fluxes estimated by our method with the flux ratios computed by METAFoR analysis techniques [MFC$^+$01, Szy95] from the data produced by FCAL. For example, the ratio of the fluxes producing serine either from glycine or glyceraldehyde 3-P was $0.36 : 0.64$, according

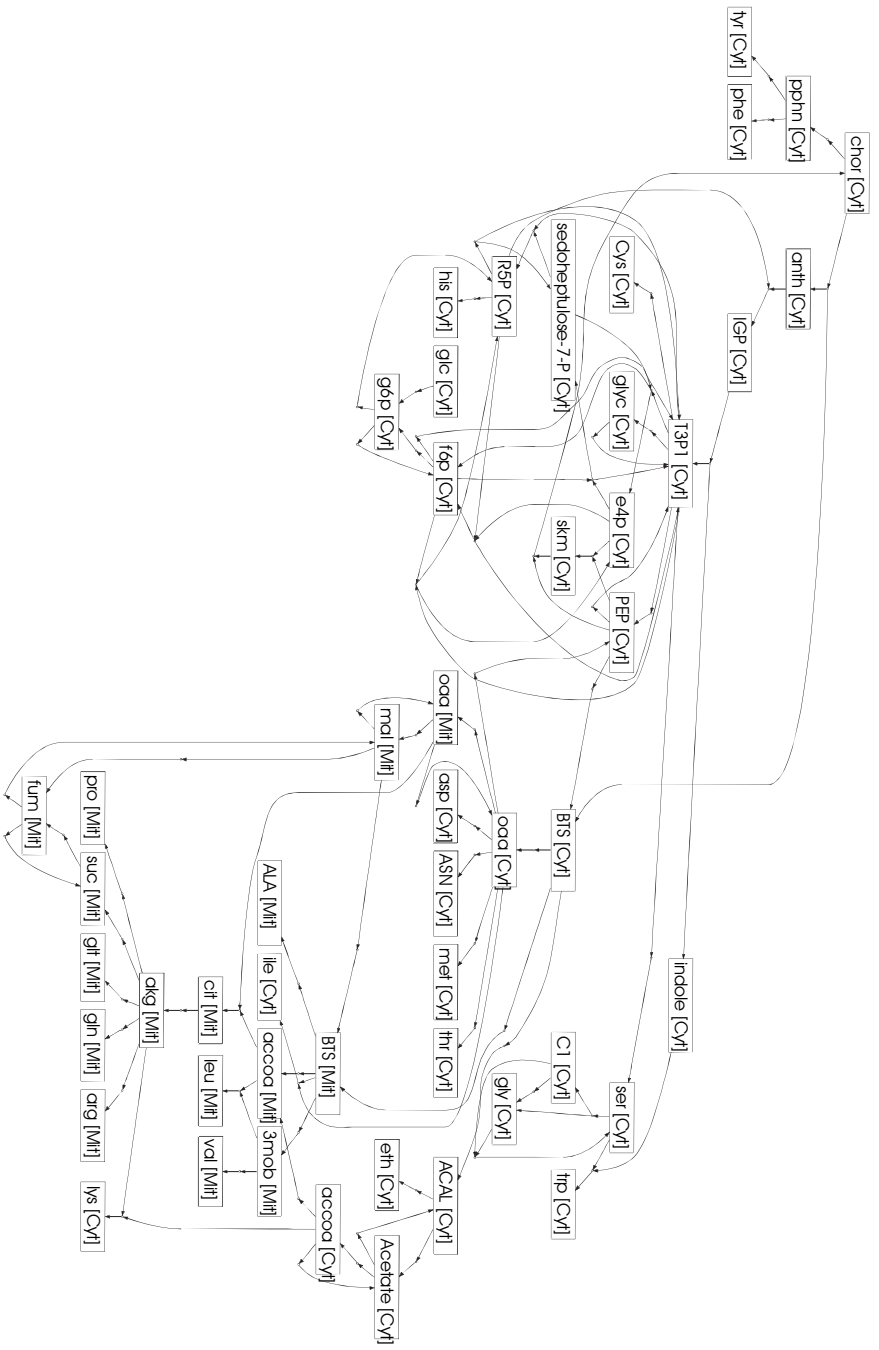


Figure 4.6: The model of the central carbon metabolism of *S. cerevisiae* used in the experiments. To reduce cluttering, external metabolites are excluded from the figure.

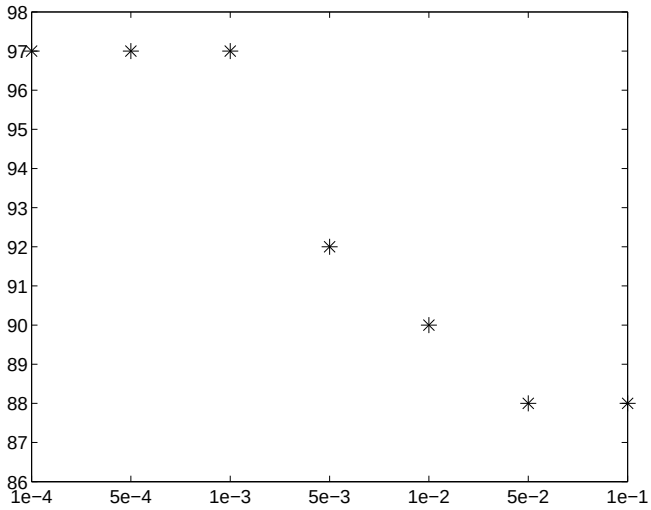


Figure 4.7: The rank of the coefficient matrix of linear equation system (4.5) constraining the fluxes (y-axis) when different cut-off values for small singular values (x-axis) were applied.

to METAFoR analysis techniques. According to our method, the same flux ratio was 0.35 : 0.65.

Chapter 5

Preprocessing MS-MS data

In this chapter we introduce computational methods to preprocess MS-MS data to a form suitable for ^{13}C metabolic flux analysis framework of Chapter 4. The methods are described in detail in Publication IV and Publication V. A reader is assumed to be familiar with the basic concepts of MS-MS introduced in Section 2.4.

Before isotopomer distributions of metabolites are measured with MS-MS, the molecules of different metabolites are separated. In the following we assume that a sample subjected to MS-MS contains only molecules of a single metabolite $\mathcal{M}$ (note that $\mathcal{M}$ refers to the whole metabolite, not only to its carbon locations M). We also assume that low level analysis of MS-MS data, such as peak detection [KO05], is already carried out. Thus MS-MS produces spectra whose peaks correspond to the relative abundances of different mass isotopomers of $\mathcal{M}$ and its different fragments $\mathcal{M}|\mathcal{F}_i$ that were cleaved from $\mathcal{M}$ during the fragmentation phase (see Section 2.4). However, in the method of direct ^{13}C metabolic flux analysis described in Chapter 4 constraints (2.8) to the isotopomer distribution of the carbon part of a metabolite are required.

The computation of constraints (2.8) to the isotopomer distribution of carbon chain M of metabolite $\mathcal{M}$ from MS-MS spectra contains the following steps:

1. Identification of fragments $\mathcal{M}|\mathcal{F}_i$. This step consists of finding the element composition of each $\mathcal{M}|\mathcal{F}_i$ and the mapping from the elements of $\mathcal{M}|\mathcal{F}_i$ to corresponding elements in $\mathcal{M}$.

2. Computation of mass isotopomer distributions $D(M)^m$ and $D(M|F_i)^m$ of carbon chains M and $M|F_i$ from the spectra describing the mass isotopomer distributions $D(\mathcal{M})^m$ and $D(\mathcal{M}|\mathcal{F}_i)^m$. Here, the effect of

naturally occurring heavy isotopes of other elements than carbon is removed from $\mathcal{M}$ and $D(M|F_i)^m$.

3. Based on $D(M)^m$ and $D(M|F_i)^m$, forming and solving a linear equation system constraining the isotopomer distribution $D(M)$ of carbon chain M.

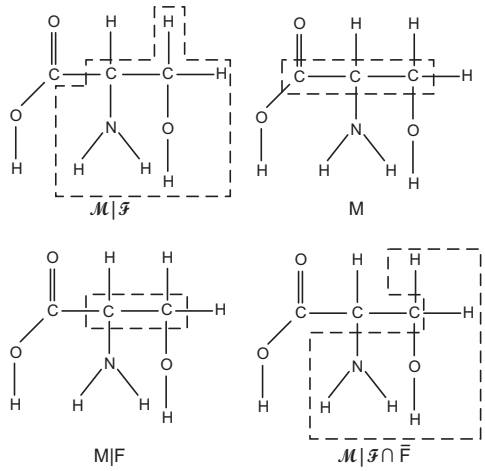


Figure 5.1: Different partitions of molecule $\mathcal{M}$: fragment $\mathcal{M}|\mathcal{F}$ (top left); carbon part M of molecule $\mathcal{M}$ (top right); carbon part $M|F$ of fragment $\mathcal{M}|\mathcal{F}$ (bottom left); non-carbon part $\mathcal{M}|\mathcal{F} \cap \overline{F}$ of fragment $\mathcal{M}|\mathcal{F}$ (bottom right).

In the next sections we describe computational methods for these steps. (As shown in Publication V, steps 2) and 3) can be merged. Here the steps are described separately for better readability.) Figure 5.1 illustrates the different partitions of molecule $\mathcal{M}$ utilized during computation.

5.1 Identification of metabolite fragments

To obtain constraints $D(M)$ from the mass isotopomer distributions $D(\mathcal{M}|\mathcal{F}_i)$ of fragments $\mathcal{M}|\mathcal{F}_i$ produced by MS-MS, we first need to know which elements of $\mathcal{M}$ belong to each $\mathcal{M}|\mathcal{F}_i$. In other words, we have to model the fragmentation of $\mathcal{M}$ in MS-MS.

The fragmentation of a molecule is a complex and stochastic process that can contain many intermediate steps. The accurate modeling of this

process is very tedious [RHO00, SHS01] and is not done in practice when fragments are identified in everyday laboratory work.

In [HRM+06] we propose an *ab initio* method for the identification of MS-MS fragments, that is based on the combinatorial analysis of the 2D structures of molecules. Shortly, we model molecule $\mathcal{M}$ as a graph, where nodes correspond to elements of $\mathcal{M}$ and edges the bonds between the elements. Furthermore, we model fragments of $\mathcal{M}$ produced in MS-MS as connected subgraphs of $\mathcal{M}$. By depth first traversal algorithm [RR00] we first generate all possible candidate fragments of the molecule whose masses correspond to the base masses of observed fragment peaks. Then we rank the candidate fragments according to the energy in the bonds that are cleaved when the fragment is produced from the molecule.

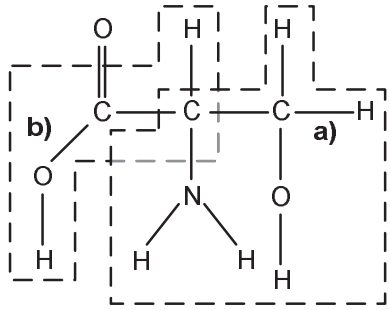


Figure 5.2: Two alternative fragments of mass 42 of serine.

Figure 5.2 depicts an example graph representing serine molecule ($C_3H_7NO_3$). Let us assume, that a peak with integer weight 42 is observed in fragment spectrum of serine. There exists many ways of cleaving a fragment with weight 42 from serine molecule, of which two are visible in the figure. In candidate fragment a) (C_2H_4N) three bonds are cleaved. The sum of the energies of cleaved bonds is 1096 J/mol and the strongest cleaved bond is $C-O$ bond of 360 kJ. In candidate fragment b) (C_2H_2O) also three bonds are cleaved. However, the sum of the energies of cleaved bonds is 1396 J/mol and cleaved bonds contain a double bond of 743 J/mol between carbon and oxygen. As we assume that strong double bonds are not likely to break, we rank candidate fragment a) ahead of candidate fragment b).

Software implementing our *ab initio* method for MS-MS fragment identification is freely available at `http://www.cs.helsinki.fi/group/sysfys/software/fragid/index.html`.

5.2 Removing the effect of natural abundance of heavy isotopes

After the identification of fragments produced by MS-MS, we are ready to compute the isotopomer distribution $D(M)$ of the carbon part M of metabolite $\mathcal{M}$, as shown in Publication IV and Publication V. As a first step of the computation, the effect of naturally occurring heavy isotopes in other elements than carbon (^2H,^{18}O, etc.) is removed from the measured mass isotopomer distributions of $\mathcal{M}$ and fragments $\mathcal{M}|\mathcal{F}_i$. This can be accomplished by a methodology introduced by Lee *et al.* [LBB91]. Let $\overline{M}$ denote the non-carbon part of $\mathcal{M}$ (thus $\mathcal{M} = M \cup \overline{M}$). Then the mass isotopomer distribution $D(M)^m$ of the carbon chain of $\mathcal{M}$ can be computed from the linear equation system, that contains an equation

$$P_{\mathcal{M}}(+l) = \sum_{h+j=l} P_{\overline{M}}(+h) P_M(+j) \tag{5.1}$$

for each observed component (that is, peak in mass spectrum) l in $D(\mathcal{M})$. In (5.1), each component of $D(\mathcal{M})$ is stated as a sum of abundances of every possible combination of distributing l extra neutrons to carbon and non-carbon parts of $\mathcal{M}$. The abundances $P_{\overline{M}}(+h)$ can be computed utilizing tables of natural abundances of isotopes (See (1) and (2) of Publication IV). Thus (5.1) is triangular and can be solved with standard techniques.

The above method can be applied also to correct mass isotopomer distributions of fragments produced by MS-MS in full scanning mode. With daughter scanning, where only a selected mass isotopomers of $\mathcal{M}$ are further fragmented, we have to be more careful. Let $\mathcal{M}|\mathcal{F} \cap \overline{F}$ denote non-carbon part of fragment $\mathcal{F}$. Heavy mass isotopomers of $\mathcal{M}$ contain more heavy natural isotopes than $\mathcal{M}$ molecules on the average do. Thus the abundance of heavy isotopes in $\mathcal{M}|\mathcal{F} \cap \overline{F}$ can not be directly computed using tables of natural abundances. In Publication IV we show how to compute the abundances of heavy natural isotopes in fragments produced by daughter ion scanning mode. By applying Bayes rule, we compute for each relevant k and l conditional probabilities $P_{\mathcal{M}|\mathcal{F}\cap\overline{F}}(+k|\mathcal{M}(+l))$ of the occurrence of k extra neutrons in non-carbon part of $\mathcal{M}|\mathcal{F}$, when only mass isotopomers $\mathcal{M}(+l)$ are fragmented ((4) of Publication IV). These conditional probabilities can then be used instead of natural abundances when daughter ion spectra are corrected by the method of Lee *et al.* [LBB91]. In Publication V we extend the method of Publication IV to handle also overlapping daughter ion spectra, as long as the base masses of the fragments are different (see Section 3.3 of Publication V).

5.3 Constraints to isotopomer distribution from MS-MS data

Mass isotopomer distributions $D(M)^m$ and $D(M|F)^m$ of carbons chains of molecule $\mathcal{M}$ and its fragments F give constraints to isotopomer distribution $D(M)$ of carbon chain of $\mathcal{M}$. From $D(M)^m$ and $D(M|F)^m$ of fragments produced by full scanning, the constraints to $D(M)$ can be stated easily. For $D(M)^m$,

$$P_M(+k) = \sum_{\sum_t b_t = k} P_M(b), \tag{5.2}$$

that is, mass isotopomer $M(+k)$ consists of those isotopomers of M that has exactly k labels. Fragments $M|F$ produced by full scanning can be dealt with analogously [CN99]:

$$P_{M|F}(+k) = \sum_{\sum_{t\in F} b_t = k} P_M(b), \tag{5.3}$$

that is, mass isotopomer $M|F(+k)$ consists of those isotopomers of M that have exactly k labels in carbons that belong to F.

With daughter ion scanning, situation is again more complicated. When only mass isotopomers $\mathcal{M}(+l)$ are fragmented, the fragment spectrum of $\mathcal{M}|\mathcal{F}$ contains only those isotopomers of M whose mass do not exceed $\mathcal{M}(+l)$. Furthermore, only a fraction of isotopomer $M(b)$ belongs to the mass isotopomer $\mathcal{M}(+l)$ and is thus fragmented. These fractions

$$i_{kjl} = \frac{P_M\left(b|\mathcal{M}(+l)\right)}{P_M(b)} \tag{5.4}$$

have to be included as coefficients to equations constraining $D(M)$:

$$P_{M|F}\left(+k|\mathcal{M}(+l)\right) = \sum_{\sum_{t\in F} b_t \leq l} i_{kjl} P_M(b). \tag{5.5}$$

Fortunately, fractions (5.4) can be precomputed (see Equation (9) of Publication IV).

Finally, all constraints (5.2), (5.3) and (5.5) can be collected to the same system and solved simultaneously to get (linear constraints to) $D(M)$ in a form compatible with(2.8).

In Publication V we rigorously formalize the computational steps required for removal of the effect of natural heavy isotopes and computation of constraints to isotopomer distribution of the carbon chain of a

metabolite. We also merge these steps to the construction of a single linear equation system. Matlab software implementing the above method for computing constraints to the isotopomer distribution of the carbon chain of a molecule from MS-MS data is freely available at `http://www.cs.helsinki.fi/group/sysfys/software/pidc/index.html`.

Chapter 6

Summary and conclusion

In this thesis we have described computational methods for ^{13}C metabolic flux analysis. The methods are based on the rigorous analysis of the combinatorics of ^{13}C labelling systems and linear algebra. The methods can be applied with all metabolic network topologies and labellings of input substrates. The manipulation of isotopomer measurements as linear subspaces facilitates the simultaneous use of measurement data produced by different measurement techniques. On the other hand, partition of metabolic fragments to equivalence classes with the help of the flow analysis techniques facilitates automatic, safe and efficient propagation of measurement data in the network thereby making it possible to constrain the fluxes more efficiently. Equivalence classes also give insight to metabolic network models with respect to ^{13}C metabolic flux analysis. They reveal redundant measurements and can explain why some fluxes cannot be identified, regardless of the analysis techniques used. Thanks to the automatic measurement propagation and generation of an equation system constraining the fluxes – as well as the wide applicability of the methods with all network topologies, substrate labellings and measurement techniques – the proposed method can be immediately applied as a new model of metabolic network is constructed, without any need for manual inspection of the properties of the network. Thus methods can be seen as generalizations and formalizations of existing methods for direct ^{13}C metabolic flux analysis that are tailored for specific measurement technologies, substrate labellings or network topologies. The proposed methods are computationally efficient.

6.1 Future work

The methods of ^{13}C metabolic flux analysis – also those presented in this thesis – are based on the assumptions given in Section 3.1. However, not all of these assumptions are trouble-free [vWVH01]. Complete metabolic networks of relative simple organisms such as *Saccharomyces cerevisiae* contain hundreds of reactions [DHP04]. The reconstruction of these networks is tedious, even if carbon atom mappings and reaction reversibility information needed for ^{13}C metabolic flux analysis is ignored. Furthermore, when models are built for ^{13}C metabolic flux analysis, a trade-off between the completeness and identifiability is faced. On the one hand, a model should contain all the reactions producing or consuming the metabolites in the model, otherwise the balance equations are not valid. On the other hand, if the model contains too many reactions, usually limited amount of measurement data cannot identify the fluxes. Thus, the correctness and the completeness of the model are not trivial objectives in ^{13}C metabolic flux analysis.

Because of the limited amount measurement data, further constraints to the fluxes are often obtained by fixing the ratios of the fluxes producing the biomass to values found from an earlier literature or by assuming that some reactions are unidirectional and that the isotopomer distributions of some metabolite pools are equivalent. These assumptions can be problematic, especially if constraints originate from different experimental conditions and different strains than used in the current experiments. Furthermore, the assumption that metabolite molecules are fully mixed in the compartments in the cell is problematic. In a mechanism called metabolic channelling the products of a reaction are transferred to the next reaction without (completely) mixing them to a common pool [KWC96, SO99]. The metabolic channelling can lead to microcompartmentation, where reactions in the same cellular compartment do not sample different isotopomers of the same metabolite from the same distribution [vWVH01]. Microcompartmentation is not usually included in the models used in ^{13}C metabolic flux analysis.

In addition to the further development of computational methods relying on the common assumptions of ^{13}C metabolic flux analysis, it might be worthwhile to develop computational tools for testing these assumptions. In this task, the fragment equivalence classes introduced in the thesis might be helpful. By definition, isotopomer distributions of fragments in the same equivalence class are equal. If competing models of the metabolic network of an organism lead to different partitions of equivalence classes, it should

be possible to come up with an experiment planning algorithm that suggests substrate labellings and metabolites to measure in such a way that only a fraction of the competing models can be consistent with the measurements (cf. [ESR+06, ZPG+03]). This task, like most other tasks of ^{13}C metabolic flux analysis, would greatly benefit from improvements in measurement technology that would allow routine measurements of intermediate metabolites [vWvDR+05].

A technical continuation to the contributions of the thesis would be the generalization of the given notion of equivalence. Currently, the equivalence relation between the fragments is derived from the dominance concept introduced in Section 4.3. However, the weak dominance defined in Section 4.3.1 is also sufficient for preserving some constraints to the isotopomer distribution. More specifically, if fragment F weakly dominates fragment E, it is guaranteed that carbons of F are transported to E as an intact fragment via all pathways, but the carbon mappings, that is, the orientation of fragments might be different in different pathways. The intactness of the fragment guarantees that labelling patterns of molecular fragments do not change. Thus, if we computed all possible joint isotopomer mappings between F and E, we would know which sums of isotopomer abundances are necessarily the same in F and E. As a safe shortcut, we could propagate mass isotopomer distributions between F and E: mass isotopomer $E(+k)$ contains all possible ways to distribute k labels to F, thus it also contains all possible images of isotopomers belonging to $F(+k)$, if F weakly dominates E.

In the ^{13}C metabolic flux analysis procedure proposed in Chapter 4, at least the very important questions of structural identifiability of the fluxes and the sensitivity of the estimated fluxes to measurement errors deserve further attention. Preprocessing methods for MS-MS data described in Publication IV and Publication V would benefit from the analysis of consistency of computed isotopomer constraints (cf. [WDW04]). Last, but certainly not least, the proposed framework for ^{13}C metabolic flux analysis should be applied (and possibly tuned) to complex, real world flux analysis tasks.

References

[ABS03] M. Araúzo-Bravo and K. Shimizu. An improved method for statistical analysis of metabolic flux analysis using isotopomer mapping matrices with analytical expressions. *Journal of Biotechnology*, 105(1–2):117–133, 2003.

[ACK+99] G. Ausiello, P. Crescenzi, V. Kann, A. Marchetti-Spaccamela, and M. Protasi. *Complexity and Approximation: Combinatorial Optimization Problems and Their Approximability Properties*. Springer, 1999.

[AHU74] A. V. Aho, J. E. Hopcroft, and J. D. Ullman. *The Design and Analysis of Computer Algorithms*. Addison Wesley, 1974.

[AHWT99] S. Alstroup, D. Harel, Lauridsen P. W., and M. Thorup. Dominators in linear time. *SIAM Journal on Computing*, 28(6):2117–2132, 1999.

[AJL+02] B. Alberts, A. Johnson, J. Lewis, M. Raff, K. Roberts, and P. Walter. *Molecular biology of the cell*. Garland Science, 4th edition, 2002.

[AKS06] M. R. Antoniewicz, J. K. Kelleher, and G. Stephanopoulos. Determination of confidence intervals of metabolic fluxes estimated from stable isotopome measurements. *Metabolic Engineering*, 8(4):324–337, 2006.

[App98] A. Appel. *Modern Compiler Implementation in Java*. Cambridge University Press, 1998.

[Ari99] M. Arita. *Automated metabolic reconstruction: theory and experiments*. PhD thesis, Tokyo University, 1999.

[Ari03] M. Arita. In silico atomic tracing of substrate-product relationships in *Escherichia coli* intermediary metabolism. *Genome Research*, 13(11):2455–2466, 2003.

[ARM] Automatic reconstruction of metabolism. http://www.metabolome.jp/index.html.

[Bea05] A. Bairoch et al. The universal protein resource (UniProt). *Nucleic Acid Research*, 33(Database issue):D154–D159, 2005.

[BKML+04] D. Benson, I. Karsch-Mizrachi, D. Lipman, J. Ostell, and D. Wheeler. Genbank: update. *Nucleic Acid Research*, 32(Database issue):D23–D26, 2004.

[BKRW98] A. L. Buchsbaum, H. Kaplan, A. Rogers, and J. R. Westbrook. A new, simpler linear-time dominators algorithm. *ACM Transactions on Programming Languages and Systems*, 20(6):1265–1296, 1998.

[BKS05] L. Blank, L. Kuepfer, and U. Sauer. Large-scale ^{13}C-flux analysis reveals mechanistic principles metabolic network robustness to null mutations in yeast. *Genome Biology*, 6(6):R49, 2005.

[BLS05] L. Blank, F. Lehmbeck, and U. Sauer. Metabolic-flux and network analysis in fourteen hemiascomycetous yeasts. *FEMS Yeast Research*, 5(6–7):545–558, 2005.

[BM03] A. P. Burgard and C. D. Maranas. Probing the performance limits of the *Escherichia coli* metabolic network subject to gene additions or deletions. *Biotechnology and Bioengineering*, 74(5):364–375, 2003.

[BSCL04] L. Boros, N. Serkova, M. Cascante, and W-N. Lee. Use of metabolic pathway flux information in targeted cancer drug design. *Drug Discovery Today: Therapeutic Strategies*, 1(4):435–443, 2004.

[BST97] H. P. J. Bonarius, G. Schmidt, and J. Tramper. Flux analysis of underdetermined metabolic networks: the quest for the missing constraints. *Trends in Biotechnology*, 15(8):308–314, 1997.

[BTS02] J Berg, J. Tymoczko, and L. Stryer. *Biochemistry*. W. H. Freeman and Company, 5th edition, 2002.

[CN99] B. Christensen and J. Nielsen. Isotopomer analysis using GC-MS. *Metabolic Engineering*, 1(4):282–290, 1999.

[CN00] B. Christensen and J. Nielsen. Metabolic network analysis of *penicillium chrysogenum* using ^{13}C-labeled glucose. *Biotechnology and Bioengineering*, 68(6):652–659, 2000.

[DBS01] M. Dauner, J. E. Bailey, and U. Sauer. Metabolic flux analysis with a comprehensive isotopomer model in *Bacillus subtilis*. *Biotechnology and Bioengineering*, 76(2):144–156, 2001.

[dGMM+00] A. A. de Graaf, M. Mahle, M. Möllney, W. Wiechert, P. Stahlmann, and H. Sahm. Determination of full ^{13}C isotopomer distributions for metabolic flux analysis using heteronuclear spin echo difference NMR spectroscopy. *Journal of Biotechnology*, 77(1):25–35, 2000.

[DHP04] N. C. Duarte, M. J. Herrgård, and B. Ø. Palsson. Reconstruction and validation of *Saccharomyces cerevisiae* iND750, a fully compartmentalized genome-scale metabolic model. *Genome Research*, 14(7):1298–1309, 2004.

[DS00] M. Dauner and U. Sauer. GC-MS analysis of amino acids rapidly provides rich information for isotopomer balancing. *Biotechnology Progress*, 16(4):642–649, 2000.

[ECP02] J. S. Edwards, M. Covert, and B. Ø. Palsson. Metabolic modelling of microbes: the flux-balance approach. *Environmental Microbiology*, 4(3):133–140, 2002.

[EDP+02] M. Emmerling, M. Dauner, A. Ponti, J. Fiaux, M. Hochuli, T. Szyperski, K. Wüthrich, J. E. Bailey, and U. Sauer. Metabolic flux responses to pyruvate kinase knockout in *escherichia coli*. *Journal of Bacteriology*, 184(1):152–164, 2002.

[EIP01] J. S. Edwards, R. U. Ibarra, and B. Ø. Palsson. In silico predictions of *Escherichia coli* metabolic capabilities are consistent with experimental data. *Nature Biotechnology*, 19(2):125–130, 2001.

[EP00] J. S. Edwards and B. Ø. Palsson. The *escherichia coli* MG1655 in silico metabolic genotype: its definition, characteristics and capabilities. *Proceedings of the National Academy of Sciences of the United States of America (PNAS)*, 97(10):5528–5533, 2000.

[ESBB98] M. B. Eisen, P. T. Spellman, P. O. Browndagger, and D. Botstein. Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences of the United States of America (PNAS)*, 95(25):14863–14868, 1998.

[ESR+06] W. Eisenreich, J. Slaghuis, Laupitz R., J. Bussemer, J. Stritzker, C. Schwarz, R. Schwarz, T. Dankekar, W. Goebel, and A. Bacher. ^{13}C isotopologue perturbation studies of listeria monocytogenes carbon metabolism and its modulation by the virulence regulator PrfA. *Proceedings of the National Academy of Sciences of the United States of America (PNAS)*, 103(7):2040–2045, 2006.

[Fea99] A. Flores et al. A protein–protein interaction map of yeast RNA polymerase III. *Proceedings of the National Academy of Sciences of the United States of America (PNAS)*, 96(14):7815–7820, 1999.

[FGS05] A. Fernie, P. Geigenberger, and M. Stitt. Flux an important, but neglected, component of functional genomics. *Current Opinion in Plant Biology*, 8(2):174–182, 2005.

[Fie02] O Fiehn. Metabolomics – the link between genotypes and phenotypes. *Plant Molecular Biology*, 48(1–2):155–171, 2002.

[FNS04] E. Fisher, Zamboni N., and U. Sauer. High-throughput metabolic flux analysis based on gas chromatography-mass spectrometry derived ^{13}C constraints. *Analytical Biochemistry*, 325(2):308–316, 2004.

[FNS05] E. Fisher, Zamboni N., and U. Sauer. FiatFlux - a software for metabolic flux analysis from ^{13}C-glucose experiments. *BMC Bioinformatics*, 6(209), 2005.

[FS03] E. Fisher and U. Sauer. Metabolic flux profiling of *Escherichia coli* mutants in central carbon metabolism using

GC-MS. *European Journal of Biochemistry*, 270(5):880–891, 2003.

[FS05] E. Fisher and U. Sauer. Large-scale *in vivo* flux analysis shows rigidity and suboptimal performance of *Bacillus subtilis* metabolism. *Nature Genetics*, 37(6):636–640, 2005.

[FTKL04] A. Fernie, R. Trethewey, A. Krotzky, and Willmitzer L. Metabolite profiling: from diagnostics to systems biology. *Nature Reviews Molecular Cell Biology*, 5(9):763–769, 2004.

[FW05] O. Frick and C. Wittmann. Characterization of the metabolic shift between oxidative and fermentative growth in *Saccharomyces cerevisiae* by comparative ^{13}C flux analysis. *Microbial Cell Factories*, 4(30), 2005.

[GCNO05] T. Grotkjr, P. Christakopoulos, J. Nielsen, and L. Olsson. Comparative metabolic network analysis of two xylose fermenting recombinant *saccharomyces cerevisiae* strains. *Metabolic Engineering*, 7(5–6):437–444, 2005.

[GMdSCN01] A. K. Gombert, M. Moreira dos Santos, B. Christensen, and J. Nielsen. 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, 2001.

[GOH+02] S. Goto, Y. Okuno, M. Hattori, T. Nishioka, and M. Kanehisa. LIGAND: database of chemical compounds and reactions in biological pathways. *Nucleic Acid Research*, 30(1):402–404, 2002.

[GWV03] H. Ge, A. Walhout, and M. Vidal. Integrating 'omic' information: a bridge between genomics and systems biology. *Trends in Genetics*, 19(10):551–559, 2003.

[GZG+05] S. Ghosh, T. Zhu, I. E. Grossmann, M. M. Ataai, and M. M. Domach. Closing the loop between feasible flux scenario identification for construct evaluation and resolution of realized fluxes via NMR. *Computers and Chemical Engineering*, 29(3):459–466, 2005.

[HDH+00] M. Huerta, G. Downing, F. Haseltine, B. Seto, and Y. Liu. http://www.bisti.nih.gov/CompuBioDef.pdf, 2000.

[Hea02] Y. Ho et al. Systematic identification of protein complexes in *saccharomyces cerevisiae* by mass spectrometry. *Nature*, 415(6868):180–183, 2002.

[Hei05] J. J. Heijnen. Approximative kinetic formats used in metabolic network modeling. *Biotechnology and Bioengineering*, 91(5):534–545, 2005.

[Hel03] M. K. Hellerstein. In vivo measurement of fluxes through metabolic pathways: the missing link in functional genomics and pharmaceutical research. In *Annual Review of Nutrition*, volume 23, pages 379–402. Annual reviews, Palo Alto, USA, 2003.

[HO93] P. C. Hansen and D. P. O'Leary. The use of the L-curve in the regularization of discrete ill-posed problems. *SIAM Journal on Scientific Computing*, 14(6):1487–1503, 1993.

[Hoo] L. Hood. `http://www.systemsbiology.org/systems_biology_in_depth`.

[HRM+06] M. Heinonen, A. Rantanen, T. Mielikäinen, E. Pitkänen, J. Kokkonen, and J. Rousu. *Ab Initio* prediction of molecular fragments from tandem mass spectrometry data. In *German Conference on Bioinformatics*, pages 40–53, 2006.

[HS96] R. Heinrich and S. Schuster. *The Regulation of Cellular Systems*. Chapman & Hall, 1996.

[IGH01] T. Ideker, T. Galsitski, and L. Hood. A new approach to decoding life: Systems biology. In *Annual Review of Genomics and Human Genetics*, pages 243–272. Annual reviews, Palo Alto, USA, 2001.

[IW03] N. Isermann and W. Wiechert. Metabolic isotopomer labeling systems. part II: structural identifibiality analysis. *Mathematical Biosciences*, 183(2):175–214, 2003.

[Kit02] H. Kitano. Computational systems biology. *Nature*, 420(14):206–210, 2002.

[KO05] M. Katajamaa and M. Orešič. Processing methods for differential analysis of LC/MS profile data. *BMC Bioinformatics*, 6(179), 2005.

[KS02] S. Klamt and S. Schuster. Calculating as many fluxes as possible in underdetermined metabolic networks. *Molecular Biology Reports*, 29(1–2):243–248, 2002.

[KS03] S. Klamt and J. Stelling. Stoichiometric analysis of metabolic networks. Tutorial et the 4^{th} International Conference on Systems Biology (ICSB 2003), http://icsb2003.molecool.wustl.edu/notes/Stelling.pdf, 2003.

[KWC96] B. N. Kholodenko, H. V. Westerhoff, and M. Cascante. Effect of channelling on the concentration of bulk-phase intermediates as cytosolic proteins become more concentrated. *Biochemical Journal*, 313(3):921–926, 1996.

[KZM+04] C. J. Krieger, P. Zhang, L. A. Mueller, A. Wang, S. Paley, M. Aranud, J. Pick, S. Y. Rhee, and P. D. Karp. MetaCyc: a multiorganism database of metabolic pathways and enzymes. *Nucleic Acids Research*, 32(Database issue):D438–442, 2004.

[Laz04] Y. Lazebnik. Can a biologist fix a radio? – or, what i learned while studying apoptosis. *Cancer Cell*, 2(3):179–182, 2004.

[LBB91] W.N. Lee, L. O. Byerley, and E. A. Bergner. Mass isotopomer analysis: theoretical and practical considerations. *Biological Mass Spectrometry*, 20(8):451–458, 1991.

[Lea04] E.S. Lander et al. Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860–921, 2004.

[LH01] H. C. Lange and J. J. Heijnen. Statistical reconciliation of the elemental and molecular biomass composition of *Saccharomyces cerevisiae*. *Biotechnology and Bioengineering*, 75(3):334–344, 2001.

[LT79] T. Lengauer and R. Tarjan. A fast algorithm for finding dominators in a flowgraph. *ACM Transactions on Programming Languages and Systems*, 1(1):121–141, 1979.

[LW00] D. J. Lockhard and E. A. Winzelar. Genomics, gene expression and DNA arrays. *Nature*, 405(6788):827–835, 2000.

[Mar98] R. E. March. An introduction to quadrupole ion trap mass spectrometry. *Journal of Mass Spectrometry*, 32(4):351–369, 1998.

[Mar01] A. Martin. General mixed integer programming: Computational issues for branch-and-cut algorithms. In Michael Jünger and Denis Naddef, editors, *Computational Combinatorial Optimization: Optimal and Provably Near-Optimal Solutions*, volume 2241 of *Lecture Notes in Computer Science*, pages 1–25. Springer, 2001.

[McL80] F. McLafferty. *Interpretation of Mass Spectra*. University Science Books, 3rd edition, 1980.

[MdGW+96] A. Marx, A. de Graaf, W. Wiechert, L. Eggeling, and H. Sahm. 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, 1996.

[MFC+01] H. Maaheimo, J. Fiaux, Z. P. Cakar, J. E. Bailey, U. Sauer, and T. Szyperski. Central carbon metabolism of *Saccharomyces cerevisiae* explored by biosynthetic fractional ^{13}C labeling of common amino acids. *European Journal of Biochemistry*, 268(8):2464–2479, 2001.

[MH03] R. Mahadevan and Schilling C. H. The effects of alternate optimal solutions in constraint-based genome-scale metabolic models. *Metabolic Engineering*, 5(4):264–276, 2003.

[MK98] P. Mendes and D. B. Kell. Non-linear optimization of biochemical pathways: applications to metabolic engineering and parameter estimation. *Bioinformatics*, 14(10):869–883, 1998.

[MMB03] C. G. Moles, P. Mendes, and J. R. Banga. Parameter estimation in biochemical pathways: A comparison of global optimization methods. *Genome Research*, 13(11):2467–2474, 2003.

[Mol04] C. Moler. *Numerical Computing with MATLAB*. Mathworks, http://www.mathworks.com/moler/, 2004.

[MWKdG99] M. Möllney, W. Wiechert, D. Kownatzki, and A. de Graaf. Bidirectional reaction steps in metabolic networks IV: Optimal design of isotopomer labeling systems. *Biotechnology and Bioengineering*, 66(2):86–103, 1999.

[MZSL98] F. McLafferty, M.-Y. Zhang, D. Stauffer, and S. Loh. Comparison of algorithms and databases for matching unknown mass spectra. *Journal of American Society for Mass Spectrometry*, 9:92–95, 1998.

[Neu98] A. Neumaier. Solving ill-conditioned and singular linear systems: A tutorial on regularization. *SIAM Review*, 40(3):636–666, 1998.

[Nie03] J. Nielsen. It is all about metabolic fluxes. *Journal of Bacteriology*, 185(24):7031–7035, 2003.

[NW06] K. Nöh and W. Wiechert. Experimental design principles for isotopically instationary ^{13}C labeling experiments. *Biotechnology and Bioengineering*, 94(2):234–251, 2006.

[Our72] E. Oura. Reactions leading to the formation of yeast cell material from glucose and ethanol. Alkon Keskuslaboratorio, Report 8078, 1972.

[PBM04] P. Pharkya, A. P. Burgard, and C. D. Maranas. Optstrain: A computational framework for redesign of microbial production systems. *Genome Research*, 14(11):2367–2376, 2004.

[PPW+03] J. A. Papin, N. D. Price, S. J. Wiback, D. A. Fell, and B. Ø. Palsson. Metabolic pathways in the post-genome era. *Trends in Biochemical Sciences*, 28(5):250–258, 2003.

[PSP03] N. D. Price, C. H. Schilling, and B. Ø. Palsson. Genome-scale microbial *in silico* models: the constraints-based approach. *Trends in Biotechnology*, 21(4):162–169, 2003.

[PTV92] W. H. Press, S. A. Teukolsky, and W. T. Vetterling. *Numerical Recipes in C: The Art of Scientific Computing*. Cambridge University Press, second edition, 1992.

[RHO00] F. Rogalewicz, Y. Hoppiliard, and G. Ohanessian. Fragmentation mechanisms of α-amino acids protonated under

electrospray ionization: a collision activation and ab initio theoretical study. *International Journal of Mass Spectrometry*, 195/196:565–590, 2000.

[RR00] G. Rücker and C. Rücker. Automatic enumeration of all connected subgraphs. *MATCH Communications in Mathematical and Computer Chemistry*, 41:145–149, 2000.

[RRU02] J. Rousu, A. Rantanen, and E. Ukkonen. Flux estimation using incomplete isotopomer information. Technical Report C-2002-55, University of Helsinki, http://www.cs.helsinki.fi/u/rousu/papers/flux_estimation.pdf, 2002.

[RSH06] R. G. Ratcliffe and Y. Shachar-Hill. Measuring multiple fluxes through plant metabolic networks. *The Plant Journal*, 45(4):490–511, 2006.

[SAN98] G. Stephanopoulos, A. Aristidou, and J. Nielsen. *Metabolic engineering: Principles and Methodologies*. Academic Press, 1998.

[Sch89] H. Schwarz. *Numerical Analysis: A Comprehensive Introduction*. John Wiley & Sons, 1989.

[SCNV97a] K. Schmidt, M. Carlsen, J. Nielsen, and J. Viladsen. Modeling isotopomer distributions in biochemical networks using isotopomer mapping matrices. *Biotechnology and Bioengineering*, 55(6):831–840, 1997.

[SCNV97b] K. Schmidt, M. Carlsen, J. Nielsen, and J. Villadsen. Modeling isotopomer distributions in biochemical networks using isotopomer mapping matrices. *Biotechnology and Bioengineering*, 55(6):831–840, 1997.

[SDCK05] D. Segr, A. DeLuna, G. Church, and R. Kishony. Modular epistasis in yeast metabolism. *Nature Genetics*, 37(1):77–83, 2005.

[SFD00] Stefan Schuster, David A. Fell, and Thomas Dandekar. A general definition of metabolic pathways useful for systematic organization and analysis of complex metabolic networks. *Nature Biotechnology*, 18(1–3):326–332, 2000.

[SGH+99] T. Szyperski, R. Glaser, M. Hochuli, J. Fiaux, U. Sauer, J. Bailey, and K. Wütrich. Bioreaction network topology and metabolic flux ratio analysis by biosynthetic fractional ^{13}C labeling and two-dimensional NMR spectrometry. *Metabolic Engineering*, 1(3):189–197, 1999.

[SHB+97] U. Sauer, V. Hatzimanikatis, J. E. Bailey, M. Hochuli, T. Szyperski, and K. Wüthrich. Metabolic fluxes in riboflavin-producing *Bacillus subtilis*. *Nature Biotechnology*, 15(5):448–452, 1997.

[SHS01] T. Shoeib, A. Hopkinson, and M. Siu. Collision-induced dissociation of the Ag^+ –proline complex: Fragmentation pathways and reaction mechanisms – a synergy between experiment and theory. *The Journal of Physical Chemistry B*, 105(749), 2001.

[SKB+02] J. Stelling, S. Klamt, K. Bettenbrock, S. Schuster, and E. D. Gilles. Metabolic network structure determines key aspects of functionality and regulation. *Nature*, 14(6912):190–193, 2002.

[SKWP02] S. Schuster, S. Klamt, W. Weckwerth, and T. Pfeiffer. Use of network analysis of metabolic systems in bioengineering. *Bioprocess and Biosystems Engineering*, 24(6):363–372, 2002.

[SMY+04] A. Sola, H. Maaheimo, K. Ylönen, P. Ferrer, and T. Szyperski. Amino acid biosynthesis and metabolic flux profiling of *Pichia pastoris*. *European Journal of Biochemistry*, 271(12):2462–2470, 2004.

[SNV99] K. Schmidt, J. Nielsen, and J. Villadsen. Quantitative analysis of metabolic fluxes in *escherichia coli*, using two-dimensional NMR spectroscopy and complete isotopomer models. *Journal of Biotechnology*, 71(1–3):175–189, 1999.

[SO99] H. O. Spivey and J. Ovádi. Substrate channeling. *Methods*, 19(2):306–321, 1999.

[SSPH99] C. H. Schilling, S. Schuster, B. Ø. Palsson, and R. Heinrich. Metabolic pathway analysis: basic concepts and scientific applications in the post-genomic era. *Biotechnology Progress*, 15(3):296–303, 1999.

[Szy95] T. Szyperski. Biosynthetically directed fractional ^{13}C-labelling of proteinogenic amino acids. *European Journal of Biochemistry*, 232(2):433–448, 1995.

[TK96] J. A. Tayek and J. Katz. Glucose production, recycling, and gluconeogenesis in normals and diabetics: a mass isotopomer [U-^{13}C]glucose study. *American Journal of Physiology – Endocrinology and Metabolism*, 270(4):E709–E717, 1996.

[Tur06] S. M. Turner. Stable isotopes, mass spectrometry, and molecular fluxes: Applications to toxicology. *Journal of Pharmacological and Toxicological Methods*, 53(1):75–85, 2006.

[VP94] A. Varma and B. Ø. Palsson. Metabolic flux balancing: basic concepts, scientific and practical use. *Nature Biotechnology*, 12(10):994–998, 1994.

[VSvD92] C. Verduyn, W. A. Scheffers, and J. P. van Dijken. Effect of benzoic acid on metabolic fluxes in yeasts: a continuous-culture study on the regulation of respiration and alcoholic fermentation. *Yeast*, 8(7):501–517, 1992.

[vWHV02] W. van Winden, J.J. Heijnen, and P. J. Verheijen. Cumulative bondomers: a new concept in flux analysis from 2d [^{13}C, ^{1}h] COSY NMR data. *Biotechnology and Bioengineering*, 80(7):731–745, 2002.

[vWHVG01] W. A. van Winden, J. J. Heijnen, P. J. Verheijen, and J. Grievink. A priori analysis of metabolic flux identifiability from ^{13}C-labeling data. *Biotechnology and Bioengineering*, 74(6):505–516, 2001.

[vWSVH01] W. A. van Winden, D. Schipper, P. J. Verheijen, and J. J. Heijnen. Innovations in generation and analysis of 2d[^{13}C, ^{1}H] COSY NMR spectra for metabolic flux analysis purposes. *Metabolic Engineering*, 3(4):322–343, 2001.

[vWvDR^{+}05] W. A. van Winden, J. J. van Dam, C. Ras, R. Kleijn, J. Vinke, W. Gulik, and J. J. Heijnen. Metabolic-flux analysis of *Saccharomyces cerevisiae* CEN.PK113-7D based on mass isotopomer measurements of ^{13}C-labeled primary metabolites. *FEMS Yeast Research*, 5(6–7):559–568, 2005.

[vWVH01] W. A. van Winden, P. J. Verheijen, and J. J. Heijnen. Possible pitfalls of flux calculations based on ^{13}C-labeling. *Metabolic Engineering*, 3(2):151–162, 2001.

[WdG97] W. Wiechert and A. de Graaf. Bidirectional reaction steps in metabolic networks: I. modeling and simulation of carbon isotope labeling experiments. *Biotechnology and Bioengineering*, 55(1):101–117, 1997.

[WDW04] S. A. Wahl, M. Dauner, and W. Wiechert. New tools for mass isotopomer data evaluation in ^{13}C flux analysis: Mass isotope correction, data consistency checking, precursor relationships. *Biotechnology and Bioenginering*, 85(3):259–268, 2004.

[WH99] C. Wittmann and E. Heinzle. Mass spectrometry for metabolic flux analysis. *Biotechnology and Bioenginering*, 62(6):739–750, 1999.

[Wie02] W. Wiechert. An introduction to ^{13}C metabolic flux analysis. *Genetic Engineering*, 24:215–238, 2002.

[WMdG01] W. Wiechert, S. Möllney, Petersen, and A. de Graaf. A universal framework for ^{13}C metabolic flux analysis. *Metabolic Engineering*, 3(3):265–283, 2001.

[WMI+99] W. Wiechert, M. Möllney, N. Isermann, M. Wurzel, and A. de Graaf. Bidirectional reaction steps in metabolic networks: III. explicit solution and analysis of isotopomer systems. *Biotechnology and Bioengineering*, 66(2):69–85, 1999.

[WMWdG96] R. Wittig, M. Möllney, W. Wiechert, and A. A. de Graaf. Interactive evaluation of NMR spectra from in vivo isotope labelling experiments. In *IFAC Symposium on Computer Applications in Biotechnology (CAB 6)*, pages 230–233. Pergamon Press, 1996.

[WSdGM97] W. Wiechert, C. Siefke, A. de Graaf, and A. Marx. Bidirectional reaction steps in metabolic networks: II. flux estimation and statistical analysis. *Biotechnology and Bioengineering*, 55(1):118–134, 1997.

[WT04] W. Wiechert and R. Takors. Validation of metabolic models: Concepts, tools, and problems. In B. N. Kholodenko

and H. V. Westerhoff, editors, *Metabolic Engineering in the Post Genomic Era*, chapter 11, pages 277–320. Horizon Bioscience, 2004.

[WvWvGH05] L. Wu, W. A. van Winden, W. M. van Gulik, and J. J. Heijnen. Application of metabolome data in functional genomics: A conceptual strategy. *Metabolic Engineering*, 7(4):302–310, 2005.

[WW01] W. Wiechert and M. Wurzel. Metabolic isotopomer labeling systems part I: global dynamic behavior. *Mathematical Biosciences*, 169(2):173–205, 2001.

[WWJ01] M. Washburn, D. Wolters, and Yates J. Large-scale analysis of the yeast proteome by multidimensional protein identification technology. *Nature Biotechnology*, 19(3):242–247, 2001.

[YWH04] T. Yang, C. Wittmann, and E. Heinzle. Metabolic network simulation using logical loop algorithm and jacobian matrix. *Metabolic Engineering*, 6(4):256–267, 2004.

[ZPG+03] T. Zhu, C. Phalakornkule, S. Ghosh, I. Grossmann, R. Koepsel, M. Ataai, and M. Domach. A metabolic network analysis & NMR experiment design tool with user interface-driven model construction for depth-first search analysis. *Metabolic Engineering*, 5(2):74–85, 2003.

Part II

TIETOJENKÄSITTELYTIETEEN LAITOS
PL 68 (Gustaf Hllstrmin katu 2 b)
00014 Helsingin yliopisto

DEPARTMENT OF COMPUTER SCIENCE
P.O. Box 68 (Gustaf Hllstrmin katu 2 b)
FIN-00014 University of Helsinki, FINLAND

JULKAISUSARJA **A**

SERIES OF PUBLICATIONS **A**

Reports may be ordered from: Kumpula Science Library, P.O. Box 64, FIN-00014 University of Helsinki, FINLAND.

A-1996-4 H. Ahonen: Generating grammars for structured documents using grammatical inference methods. 107 pp. (Ph.D. thesis).

A-1996-5 H. Toivonen: Discovery of frequent patterns in large data collections. 116 pp. (Ph.D. thesis).

A-1997-1 H. Tirri: Plausible prediction by Bayesian inference. 158 pp. (Ph.D. thesis).

A-1997-2 G. Lindén: Structured document transformations. 122 pp. (Ph.D. thesis).

A-1997-3 M. Nykänen: Querying string databases with modal logic. 150 pp. (Ph.D. thesis).

A-1997-4 E. Sutinen, J. Tarhio, S.-P. Lahtinen, A.-P. Tuovinen, E. Rautama & V. Meisalo: Eliot – an algorithm animation environment. 49 pp.

A-1998-1 G. Lindén & M. Tienari (eds.): Computer Science at the University of Helsinki 1998. 112 pp.

A-1998-2 L. Kutvonen: Trading services in open distributed environments. 231 + 6 pp. (Ph.D. thesis).

A-1998-3 E. Sutinen: Approximate pattern matching with the q-gram family. 116 pp. (Ph.D. thesis).

A-1999-1 M. Klemettinen: A knowledge discovery methodology for telecommunication network alarm databases. 137 pp. (Ph.D. thesis).

A-1999-2 J. Puustjärvi: Transactional workflows. 104 pp. (Ph.D. thesis).

A-1999-3 G. Lindén & E. Ukkonen (eds.): Department of Computer Science: annual report 1998. 55 pp.

A-1999-4 J. Kärkkäinen: Repetition-based text indexes. 106 pp. (Ph.D. thesis).

A-2000-1 P. Moen: Attribute, event sequence, and event type similarity notions for data mining. 190+9 pp. (Ph.D. thesis).

A-2000-2 B. Heikkinen: Generalization of document structures and document assembly. 179 pp. (Ph.D. thesis).

A-2000-3 P. Kähkipuro: Performance modeling framework for CORBA based distributed systems. 151+15 pp. (Ph.D. thesis).

A-2000-4 K. Lemström: String matching techniques for music retrieval. 56+56 pp. (Ph.D.Thesis).

A-2000-5 T. Karvi: Partially defined Lotos specifications and their refinement relations. 157 pp. (Ph.D.Thesis).

A-2001-1 J. Rousu: Efficient range partitioning in classification learning. 68+74 pp. (Ph.D. thesis)

A-2001-2 M. Salmenkivi: Computational methods for intensity models. 145 pp. (Ph.D. thesis)

A-2001-3 K. Fredriksson: Rotation invariant template matching. 138 pp. (Ph.D. thesis)

A-2002-1 A.-P. Tuovinen: Object-oriented engineering of visual languages. 185 pp. (Ph.D. thesis)

A-2002-2 V. Ollikainen: Simulation techniques for disease gene localization in isolated populations. 149+5 pp. (Ph.D. thesis)

A-2002-3 J. Vilo: Discovery from biosequences. 149 pp. (Ph.D. thesis)

A-2003-1 J. Lindström: Optimistic concurrency control methods for real-time database systems. 111 pp. (Ph.D. thesis)

A-2003-2 H. Helin: Supporting nomadic agent-based applications in the FIPA agent architecture. 200+17 pp. (Ph.D. thesis)

A-2003-3 S. Campadello: Middleware infrastructure for distributed mobile applications. 164 pp. (Ph.D. thesis)

A-2003-4 J. Taina: Design and analysis of a distributed database architecture for IN/GSM data. 130 pp. (Ph.D. thesis)

A-2003-5 J. Kurhila: Considering individual differences in computer-supported special and elementary education. 135 pp. (Ph.D. thesis)

A-2003-6 V. Mäkinen: Parameterized approximate string matching and local-similarity-based point-pattern matching. 144 pp. (Ph.D. thesis)

A-2003-7 M. Luukkainen: A process algebraic reduction strategy for automata theoretic verification of untimed and timed concurrent systems. 141 pp. (Ph.D. thesis)

A-2003-8 J. Manner: Provision of quality of service in IP-based mobile access networks. 191 pp. (Ph.D. thesis)

A-2004-1 M. Koivisto: Sum-product algorithms for the analysis of genetic risks. 155 pp. (Ph.D. thesis)

A-2004-2 A. Gurtov: Efficient data transport in wireless overlay networks. 141 pp. (Ph.D. thesis)

A-2004-3 K. Vasko: Computational methods and models for paleoecology. 176 pp. (Ph.D. thesis)

A-2004-4 P. Sevon: Algorithms for Association-Based Gene Mapping. 101 pp. (Ph.D. thesis)

A-2004-5 J. Viljamaa: Applying Formal Concept Analysis to Extract Framework Reuse Interface Specifications from Source Code. 206 pp. (Ph.D. thesis)

A-2004-6 J. Ravantti: Computational Methods for Reconstructing Macromolecular Complexes from Cryo-Electron Microscopy Images. 100 pp. (Ph.D. thesis)

A-2004-7 M. Kääriäinen: Learning Small Trees and Graphs that Generalize. 45+49 pp. (Ph.D. thesis)

A-2004-8 T. Kivioja: Computational Tools for a Novel Transcriptional Profiling Method. 98 pp. (Ph.D. thesis)

A-2004-9 H. Tamm: On Minimality and Size Reduction of One-Tape and Multitape Finite Automata. 80 pp. (Ph.D. thesis)

A-2005-1 T. Mielikäinen: Summarization Techniques for Pattern Collections in Data Mining. 201 pp. (Ph.D. thesis)

A-2005-2 A. Doucet: Advanced Document Description, a Sequential Approach. 161 pp. (Ph.D. thesis).

A-2006-1 A. Viljamaa: Specifying Reuse Interfaces for Task-Oriented Framework Specialization. 285 pp. (Ph.D. thesis).

A-2006-2 S. Tarkoma: Efficient Content-based Routing, Mobility-aware Topologies, and Temporal Subspace Matching. 198 pp. (Ph.D. thesis).

A-2006-3 M. Lehtonen: Indexing Heterogeneous XML for Full-Text Search. 185+3 pp. (Ph.D. thesis).