Network inference from time-resolved metabolomics data
Hendrickx, D.M.

Citation for published version (APA):
Network inference from time-resolved metabolomics data

Diana M. Hendrickx


Network inference from time-resolved metabolomics data

Donderdag 18 april 2013
om 10.00 uur

Agnietenkapel op de
Oudezijds Voorburgwal 231
in Amsterdam

Na afloop van de ceremonie
is er een receptie.

Vervolgens is er na
afloop van de receptie
van 12:30 tot 14:00
een besloten lunch.

Vanaf 19:30 is iedereen
uitgenodigd om de promotie
tevieren. Dit feest is voor
iedereen open.

Paranimfen
Daniel Vis
Johan Andriessen

Diana Hendrickx
Vlietstraat 11
2627 Schelle
België
+32 (0)473 77 38 40
diana.hendrickx@hotmail.com

Uitnodiging
voor het bijwonen van
de openbare verdediging
van het proefschrift
Network inference from time-resolved metabolomics data

ACADEMISCH PROEFSCHRIFT

der verkrijging van de graad van doctor
aan de Universiteit van Amsterdam
op gezag van de Rector Magnificus
prof. dr. D.C. van den Boom
ten overstaan van een door het college voor promoties
ingestelde commissie,
in het openbaar te verdedigen in de Agnietenkapel
op donderdag 18 april 2013, te 10:00 uur

door

Diana Mathilda Hendrickx

geboren te Borgerhout, België
Promotiecommissie:

Promotor:

- prof. dr. A.K. Smilde

Copromotores:

- dr. ir. H.C.J. Hoefsloot
- dr. M.M.W.B. Hendriks

Overige leden:

- prof. dr. A.H.C. van Kampen
- prof. dr. F.J. Bruggeman
- prof. dr. B.M. Bakker
- prof. dr. M.J. Teixeira de Mattos
- prof. dr. K.J. Hellingwerf
- dr. M.W.E.M. van Tilborg

Faculteit der Natuurwetenschappen, Wiskunde en Informatica

The research reported in this thesis was carried out at the Swammerdam Institute for Life Sciences, Faculty of Science, Universiteit van Amsterdam. The project was financed by the Netherlands Metabolomics Centre (NMC), which is part of the Netherlands Genomics Initiative - Netherlands Organization for Scientific Research.
# Contents

1 Introduction ................................. 1

1.1 Time-resolved metabolomics data ............ 3

1.1.1 Measurements in metabolomics .......... 3

1.1.2 Sampling methods for time-resolved microbial metabolomics .... 5

1.1.3 Analytical techniques for time-resolved microbial metabolomics .... 6

1.1.4 Labeling experiments ................. 6

1.2 Metabolic network inference .............. 7

1.2.1 What is metabolic network inference? ... 7

1.2.2 Methods for metabolic network inference ... 8

1.3 Challenges in metabolic network inference ... 12

1.3.1 Estimating the topology and directionality from time-resolved metabolomics data ... 12

1.3.2 Incorporating pathway information ...... 12

1.3.3 Interpretation of correlations .......... 13

1.3.4 Combining experimental data with stoichiometric models .... 13

1.4 Scope and outline of the thesis .......... 14

Acknowledgments ............................ 14

2 Reverse engineering of metabolic networks, a critical assessment 15

2.1 Introduction .............................. 16
2.2 Methods .................................................. 23
  2.2.1 Similarity measures ................................. 24
  2.2.2 Time-lagged correlations ............................ 24
  2.2.3 The penalized Jacobian method ..................... 25
  2.2.4 The zero slopes method ............................. 27
  2.2.5 Method performance ................................. 28
  2.2.6 Effect of noise on network inference ............... 29
  2.2.7 Simulations .......................................... 30
2.3 Results and discussion ................................. 31
  2.3.1 Results .............................................. 31
  2.3.2 Discussion .......................................... 35
2.4 Conclusion .............................................. 37
Appendix A: Calculation of n-th order partial Pearson correla-
   tions ...................................................... 39
Appendix B: Calculation of the time-lagged correlation matrix 39
Appendix C: Algorithm to calculate the Jacobian matrix ........ 40
Appendix D: The vertex-edge incidence matrix ................. 43
Acknowledgments ........................................... 45

3 Global test for metabolic pathway differences between condi-
   tions ...................................................... 47
  3.1 Introduction ............................................ 48
  3.2 Materials and methods ................................. 50
    3.2.1 Escherichia coli data set ......................... 50
    3.2.2 Saccharomyces cerevisiae data set ............... 51
    3.2.3 Data pre-treatment ................................ 53
    3.2.4 Goeman’s global test .............................. 53
    3.2.5 Computational tools ............................... 57
  3.3 Results and discussion ................................ 58
    3.3.1 Results ............................................ 58
    3.3.2 Discussion ......................................... 61
  3.4 Conclusion ............................................ 68
Acknowledgments ........................................... 68
Supplementary Data ......................................... 68

iv
Chapter 1

Introduction

Metabolism is the whole of all chemical processes in an organism that enable growth, reproduction and adaptation to the environment. Metabolic processes can be divided in degradation processes (catabolism) and biosynthesis (anabolism) [93]. The intermediates of metabolism, metabolites, are no separate entities, but are organized in metabolic pathways, which are part of a large network. Each step in a metabolic network is catalyzed by one or more enzymes. The flux through a pathway is regulated by (genetic) metabolic control mechanisms (see Table 1.1).

Unraveling the functioning of metabolic pathways is an important goal of systems biology, because it contributes to understanding biological processes in the cell. Poorly understood properties of the cell are cellular decision-making and robustness [7, 61, 90]. Cellular decision-making systems are mechanisms that make the cell adapt effectively to changing environments [218]. Robustness is the maintenance of certain properties for survival [188].

Cellular decisions are made at the level of biochemical networks [7]. Therefore, a first step in understanding cellular decision-making is knowledge of structural and kinetic properties of biochemical networks. This can be accomplished by network inference methods.
Introduction

Table 1.1: Overview of metabolic control mechanisms.

<table>
<thead>
<tr>
<th>mechanism</th>
<th>definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>enzyme induction</td>
<td>increased enzyme synthesis in the presence or absence of a certain metabolite [73, 126]</td>
</tr>
<tr>
<td>enzyme repression</td>
<td>decreased enzyme synthesis in the presence or absence of a certain metabolite [73, 126]</td>
</tr>
<tr>
<td>substrate-level control</td>
<td>high levels of product inhibit the substrate to react [126]</td>
</tr>
<tr>
<td>feedback control</td>
<td>cell controls generation of a product through activation (positive control) or inhibition (negative control) of an earlier reaction in the pathway [93, 196, 126] (see Figure 1.1 a and b)</td>
</tr>
<tr>
<td>feed forward control</td>
<td>a metabolite activates or inhibits a further step in the pathway [196, 126] (see Figure 1.1 c and d)</td>
</tr>
</tbody>
</table>

Figure 1.1: Metabolic control mechanisms a) positive feedback control b) negative feedback control c) positive feedforward control d) negative feed forward control
Once the structure and directionality of a metabolic pathway are known, comparative analysis of metabolite correlations under different conditions can reveal additional information about the pathway [189]. The availability of resources determines how pathways adjust to their environment [90]. Pathway statistics can be used to elucidate which pathways change between two conditions. Cells are evolved towards an optimal response to changes in their environment [90, 111]. Therefore, it is important to study what the cell has optimized for a better understanding of cellular decision-making and robustness [90]. This can be achieved by methods based on optimization theory, like flux balance analysis (FBA) [147, 175]. This thesis focuses on metabolic network inference from time-resolved microbial metabolomics data. Time series describe the dynamic response of the cell to a perturbation and therefore provide more information than stationary data [13]. Using time series improves inference of causal relationships, network reconstruction and parameter estimation [13, 49, 182].

1.1 Time-resolved metabolomics data

1.1.1 Measurements in metabolomics

In this thesis, microbial metabolite profiling datasets are used to illustrate metabolic network inference methods. In metabolomics analysis, one can distinguish between semi-quantitative and quantitative measurements (see Figure 1.2). Values of semi-quantitative measurements are peak areas (relative concentrations) and can be used to study the qualitative behavior of metabolites [236]. Semi-quantitative measurement of intracellular and extracellular metabolites is called metabolic fingerprinting and metabolic footprinting respectively [124]. Quantitative measurements are metabolite concentrations expressed in chemical units (moles per gram dry weight or moles per liter) [236]. Target analysis is quantitative measurement of one or several (internal or external) metabolites of interest [222]. Metabolite profiling is quantifying preselected groups of metabolites belonging to the same pathway or with similar
chemical properties (e.g. lipids) [145].

Figure 1.2: Types of measurements in metabolomics.
1.1.2 Sampling methods for time-resolved microbial metabolomics

Figure 1.3: Operation modes of a bioreactor: a) batch b) fed-batch c) continuous. Adapted from Mashego et al [124].
Introduction

Samples for microorganisms are taken from a bioreactor. Bioreactors can operate in three different modes (see Figure 1.3): batch (no inflow, no outflow), fed-batch (inflow, no outflow), continuous (chemostat, inflow and outflow) [124].

Fast sampling devices sample a small amount of metabolites (mostly only central pathways) on a second or sub second time scale [210]. Larger groups of metabolites (tens or hundreds) are sampled with slow devices on a time scale of hours [94].

Currently, fast sampling devices only exist for microorganisms. Human, mammals and plants are sampled on the minutes or hours scale [12, 97, 8].

1.1.3 Analytical techniques for time-resolved microbial metabolomics

Analytical techniques for microbial metabolomics mostly consist of a chromatographic method for separation of metabolites, followed by mass detection using mass spectrometry (MS). In gas chromatography (GC), the analytes are separated by their physical properties. In liquid chromatography (LC), the separation is based on chemical properties [224].

1.1.4 Labeling experiments

$^{13}$C-labeled metabolite data are used to calculate intracellular fluxes, because these fluxes can not be measured directly [235]. In $^{13}$C-labeling experiments, medium substrates are labeled with $^{13}$C [212]. The labeled carbon atoms propagate through the metabolic pathways [228]. The different labeling states of the metabolites, isotopomers, can then be measured with GC-MS or LS-MS [144]. A metabolite with $n$ carbon atoms has $2^n$ isotopomers (each carbon atom can be labeled or unlabeled) [228]. For microorganisms, isotopomers can be measured on a second scale [143, 220]. Fluxes are calculated from the labeling data and measurements of external fluxes. This method is called $^{13}$C metabolic flux analysis ($^{13}$C MFA) [228].
1.2 Metabolic network inference

1.2.1 What is metabolic network inference?

Metabolic network inference is the extraction of metabolic network information from experimental data by means of a mathematical framework [196].

Metabolic pathways can be studied on different levels [190], with increasing amount of detail. The most basic level is structure identification, which consists of determining the topology of the network (see Figure 1.4a)). An edge is drawn between two metabolites if the one is converted in the other by a chemical reaction. A second way of examining a pathway is studying the stoichiometry, the amount of substrates and products involved in the reactions (see Figure 1.4b)). Thermodynamical properties of the reactions can be studied to determine the directionality [150] (see Figure 1.4c)). Finally, rate laws can be formulated and parameters be estimated, which results in a detailed kinetic model [201, 30] (see Figure 1.4d)).

Figure 1.4: Different levels of studying metabolic pathways. a) topology b) stoichiometry c) directionality d) kinetics
Inference of metabolic networks can serve different purposes, like understanding cellular functioning and generation of hypotheses [196].

1.2.2 Methods for metabolic network inference

1.2.2.1 Bottom-up and top-down approach

There are two classical approaches to network inference: bottom-up and top-down. The bottom-up or forward approach uses available knowledge on kinetic or chemical properties of the network, obtained from the literature or databases [211, 17]. This knowledge is combined to obtain large-scale models [187]. The top-down approach, also called reverse engineering, infers network properties from experimental data [17]. Recently, also middle-out approaches are applied that combine bottom-up and top-down inference [95].

1.2.2.2 Methods for determining network structure

Different mathematical and statistical top-down methods are available for determining the topology and directionality of a metabolic network. Methods based on association measures include time-lagged correlation [5], partial Pearson correlation and mutual information [27]. A second category of methods are probabilistic approaches, like Bayesian networks [45]. Other approaches are based on linear approximations of non-linear reaction models (e.g. Jacobian method) [37].

1.2.2.3 Comparing pathways between different conditions

Pathways can be compared between conditions with pathway statistics, which provide a manner to study pathways as a whole. Pathway statistics originate from microarray studies [48] and are currently extended to metabolomics [232, 29, 85]. They are based on the idea that genes and metabolites change in a coordinated way [48, 232]. Studying pathways instead of single genes or metabolites has several advantages. Subtle coordinated changes can be discovered that cannot
be detected with tests for individual genes or metabolites [232]. Furthermore, comparative studies are facilitated because the number of hypotheses that has to be tested is reduced [48].

Two different types of pathway statistics can be distinguished. Competitive tests compare a pathway with the rest of the genes or metabolites in the dataset. Self-contained tests examine if a pathway is different between two phenotypes or conditions [48].

1.2.2.4 Association networks

Association networks or relevance networks connect metabolites based on their similarity, which is characterized by a similarity measure. Metabolites are connected if the calculated similarity measure is above a certain threshold. Frequently used similarity measures are Pearson correlation, Spearman correlation and mutual information [27].

Associations in a relevance network are not necessary metabolic reactions [21]. They are the result of the combination of all reactions and regulatory interactions in the network [189].

Correlations provide information about the regulation of the underlying pathways [192]. High positive correlation of a metabolite pair can point to rapid equilibrium or dominance of an enzyme, while high negative correlation can indicate the presence of a conserved moiety [21].

Comparing correlation networks between different conditions can provide information about the invariant features of metabolic pathways, changes in regulation and the existence of multiple steady states [189]. Correlations preserved among different conditions can point to rapid equilibrium. Reversed correlations between conditions can indicate a change in regulation or the existence of multiple steady states [189].

For the reasons mentioned above, studies on association networks are equally important as studies on metabolic reaction networks.

Metabolic reaction networks and association networks provide complementary information about a metabolic pathway (network structure and regulation respectively). Therefore, combined studies of reaction and association networks provide more information than studying each type of network separately.
1.2.2.5 **Kinetic models**

Kinetic models describe metabolic networks with non-linear differential equations [196]. They are used for determining the steady-state(s) of the system, simulating time-courses and studying metabolic control [162]. Detailed information about rate laws and kinetic parameters is required for building kinetic models [119]. When the exact form of the rate laws is unknown, approximate rate laws (e.g. S-systems) can be used [196].

1.2.2.6 **Stoichiometric models**

Often, there is insufficient experimental data to estimate the parameters in a kinetic model. Stoichiometric models were developed to avoid the difficulties with kinetic models [112]. Figure 1.5 gives an overview of current methodologies in stoichiometric modeling.

![Figure 1.5: Current methodologies in stoichiometric modeling.](image-url)
Stoichiometric models make use of mass balances \(\frac{dC}{dt} = S \cdot v\), where \(S\) is the stoichiometric matrix, \(v\) the vector of reaction rates and \(dC/dt\) the time derivatives of the metabolite concentrations [162]. Stoichiometric models can be used to elucidate the systemic properties of metabolism or to determine reaction rates [112]. Systemic properties are analyzed under steady state conditions, where \(S \cdot v = 0\). Concepts often used for system analysis are elementary modes (EM) and extreme pathways (EP). Both EM and EP define all possible routes from a substrate to a product. EP assumes that all reactions are irreversible, while EM allow for reversibility [93]. Stoichiometric models for determining reaction rates can be divided in models for reaction rate estimation and predictive models.

Steady state fluxes can be estimated from mass balances and external flux measurements by metabolic flux analysis (MFA) [112]. Dynamic MFA is an extension of MFA for estimating reaction rate profiles over time [110]. Often, not enough external fluxes can be measured to make the system of mass balances determined. \(^{13}\)C MFA (see 1.1.4.) overcomes this problem [112], because measuring isotopomers instead of metabolites reduces the degrees of freedom.

Flux balance analysis (FBA) is a stoichiometric modeling approach that predicts the steady state flux distribution based on an optimality hypothesis, which describes the biological goal of the organism under a given condition [112]. The hypothesis is formulated as an objective function, which is minimized or maximized, given certain constraints [147]. The constraints are the mass balances and additional inequality constraints on the reaction rates, thermodynamics and regulation [112]. The objective function is a linear combination of the fluxes [147]. The resulting optimization problem is solved for the fluxes. The result of the FBA is a prediction of the flux distribution that will occur under the conditions determined by the constraints [112]. Often, there are alternative solutions that reach the optimum for the objective function, given the constraints. The range of optimal solutions can be studied with flux variability analysis (FVA) [121].

Dynamic FBA (DFBA) is an extension of FBA that accounts for dy-
namic changes in cellular behavior [120]. In DFBA, the mass balance constraints are differential equation constraints instead of linear constraints [196]. DFBA approaches can be divided in two groups [116]: static optimization approach (SOA) and dynamic optimization approach (DOA). In the SOA approach, the time period is divided in intervals and an optimization problem is solved at the starting point of each time interval. The DOA approach solves a single optimization problem for the entire time course [120].

1.3 Challenges in metabolic network inference

1.3.1 Estimating the topology and directionality from time-resolved metabolomics data

When estimating the topology and directionality of a metabolic network, one has to deal with several issues. In experimental data, the number of samples is often much lower than the number of metabolites in the network ("curse-of-dimensionality") [155]. Because of the curse-of-dimensionality problem, different network topologies can match with experimental observations [87]. Furthermore, experimental data have a high level of noise [196]. Therefore, it is important to find out how much noise is allowed for a good performance of network inference methods. It is also crucial to know which type of data are required for different network inference methods [196]. One needs to know what kind of perturbations are necessary, how many replicates are required, how frequent samples have to be taken and how long the time series has to be [107].

1.3.2 Incorporating pathway information

In metabolomics, there exist various univariate and multivariate statistical methods for finding significant differences under changing conditions [138, 69]. However, these methods treat the metabolites as separate units and do not take into account that metabolites are organized in pathways
A challenge for metabolomics is to explore how information about pathway structure can be incorporated into statistical methods.

### 1.3.3 Interpretation of correlations

Many statistical methods for metabolomics make use of correlations or covariances. Examples are principal component analysis (PCA) [77], canonical correlation analysis (CCA) [233] and individual differences scaling (INDSCAL) [75]. It is important to understand what these correlations mean biologically. Previous correlation studies focus on steady state data [192, 21, 189]. However, metabolite levels change dynamically in response to perturbations [87]. Correlation analysis can provide biological information additional to the information provided by steady state correlation analysis. Extracting information from correlations is a challenging task because there is no direct relationship between a correlation network and the underlying pathway [192].

### 1.3.4 Combining experimental data with stoichiometric models

A disadvantage of stoichiometric models is that they often result in a large solution space [121]. When also dynamics are included, the methods also become mathematically complex because differential equation constraints are involved. The mathematical complexity makes them less suitable for studying larger systems [120]. Integration of experimental data into stoichiometric models can reduce the solution space [159]. Examples of combining experimental data with steady state flux balance analysis are rFBA and IOMA. In rFBA, transcriptional regulation is integrated into flux balance analysis [36]. IOMA combines quantitative proteomics and metabolomics data with flux balance analysis [234]. Similar mathematical methods also have to be developed for dynamic flux balance analysis, in order to reduce both the solution space and the mathematical complexity due to differential equation constraints.
1.4 Scope and outline of the thesis

This thesis focuses on the inference of metabolic network properties from time-resolved metabolite concentration data. Each chapter addresses one of the challenges described in paragraph 1.3.

Chapter 2 presents a study about the feasibility of estimating the topology and directionality of metabolic networks from time-resolved metabolomics data.

Chapter 3 deals with incorporating pathway information in studies that compare different conditions. The extension of a pathway-based method (Goeman’s global test) from gene expression analysis to metabolomics is explained in detail.

Chapter 4 focuses on extracting network information from correlations in time-resolved metabolomics data. Information about the pathway structure is combined with correlation analysis to infer regulation mechanisms responsible for changes in the distribution of reaction rates across conditions.

Chapter 5 addresses the integration of time-resolved metabolomics data into dynamic flux balance analysis (DFBA) with the aim to reduce both the solution space and the mathematical complexity of standard DFBA. Finally, some suggestions for future research are described in chapter 6.

Acknowledgments

This project was financed by the Netherlands Metabolomics Centre (NMC), which is part of the Netherlands Genomics Initiative - Netherlands Organization for Scientific Research.
Chapter 2

Reverse engineering of metabolic networks, a critical assessment

Inferring metabolic networks from metabolite concentration data is a central topic in systems biology. Mathematical techniques to extract information about the network from data have been proposed in the literature. This chapter presents a critical assessment of the feasibility of reverse engineering of metabolic networks, illustrated with a selection of methods. Appropriate data are simulated to study the performance of four representative methods. An overview of sampling and measurement methods currently in use for generating time-resolved metabolomics data is given and contrasted with the needs of the discussed reverse engineering methods. The results of this assessment show that if full inference of a real-world metabolic network is the goal there is a large discrepancy between the requirements of reverse engineering of metabolic

---

networks and contemporary measurement practice. Recommendations for improved time-resolved experimental designs are given.

2.1 Introduction

Reverse engineering of biological networks is an important topic in systems biology. Gene regulatory networks, protein-protein interaction networks and metabolic networks all have their own characteristics and difficulties. This chapter focuses on inference of metabolic networks from (time-resolved) metabolomics data. Metabolomics data can contain a wealth of information about the biochemical reactions and interactions in the cell of an organism [189]. Metabolic network inference can help to elucidate these mechanisms [27] and is therefore gaining interest in different scientific disciplines, such as microbiology, plant biology and biomedical sciences. Moreover, the metabolome is the functional genomics level closest to the cell’s phenotype [69] and thus improves our understanding of the functioning of cellular systems [27, 223]. In biomedical sciences, metabolic networks can be used for distinguishing normal from abnormal cell phenotypes. This is important for a better comprehension of diseases and can serve as a starting point for the discovery of new drugs and therapeutic methods [69].

Metabolic network inference consists of estimating one or more of four characteristics of the network: the topology (interactions between metabolites), the directionality (direction of the interactions (arrows) in the network), the stoichiometry (the number of molecules involved in each reaction) and the kinetics (flux rates). This study focuses on the connectivity of the network, that is, the first two characteristics.

A number of network inference methods are available for estimating connectivity. Some methods are based on correlations between metabolites [189, 5, 21, 135] and use steady state as well as time series [37] data. Other approaches include probabilistic methods (Bayesian networks) [45] and non-linear reaction models [37]. For each of these methods examples exist that support the claim that they perform well [37, 6, 166, 190, 213]. Most of these examples concern small networks of four to six metabolites [6, 166, 190] or networks with first order kinetics only [213]. We critically
assess the quality of reverse engineering methods on larger, non-linear systems, where network inference becomes more complicated: the glycolytic pathway of *S. cerevisiae* with 13 metabolites and the central carbon metabolism of *E. coli* with 18 metabolites. First, the number of possible connections increases quadratically with the number of metabolites, which makes the inference task computationally more complex. Secondly, the spread of time constants is usually increased in larger networks. Hence, the frequency of the measurements should increase with the size of the network and measurements over a longer time period are needed to study the slow interactions. Thirdly, our model systems include non-linear terms and higher-order kinetics.

As a reference, a literature survey regarding sampling and measurement methods in use for time-resolved metabolomics data was performed (Table 2.1). Sampling in microorganisms is, depending on the method, possible on a small time-scale, but published studies are restricted to a relatively limited number of compounds. With the fastest techniques, samples can immediately be taken from the bioreactor [199] on a subsecond time-scale [20, 40, 106, 171, 169, 199, 210]. Rapid sampling devices make use of a micro-valve [20], connected by a capillary to the reactor and the sampling tube, which is filled with cold quenching fluid (e.g. methanol) [106]. Until now, sampling on (sub)second scale has not been reported for human, mammals and plants, where repeated sampling is usually done on time-scales of minutes or hours. Measurement errors are in the order of 5-25 % (RSD).

Next, to illustrate the complexity of network recovery, simulated data are used. Using such data has several advantages: the complete underlying metabolic network is known, the sampling frequency can be adjusted and the noise level can be controlled [131]. To perform the critical assessment of reverse engineering methods, four representative methods were chosen; each approach needing data extracted on a different time-scale and another type of perturbation experiments. An overview is given in Table 2.2. First, we wanted to include a method for static data, collected at different steady states of the system. For this purpose, a method studied by Çakir and coworkers [27] that uses partial Pearson correlations (also called a Graphical Gaussian Model) is discussed. Secondly, a correlation
Reverse engineering of metabolic networks, a critical assessment

based method for time series is studied, using time-lagged correlations, as proposed by Arkin and Ross (1995) [5]. Thirdly, a model-based approach for dynamic data is discussed: the Jacobian method, presented earlier in several papers [191, 37, 172]. This method was combined with modern penalty methods to induce sparsity in the Jacobian [47, 173]. Finally, we examined the possibility to use calculations of initial slopes of dynamic concentration profiles, suggested in the literature by Crampin and coworkers [37] because of its conceptual simplicity. An overview of the properties of the four methods is given in Table 2.2.

The results of the critical evaluation based on the simulations are presented and confronted with the currently available time-resolved metabolomics data. Based on that, we give limitations of the current reverse engineering approaches and recommendations for experimental designs of time-resolved metabolomics experiments, aimed at metabolic network inference.

Table 2.1: Methods for time-resolved metabolomics measurements.

<table>
<thead>
<tr>
<th>Sampling method</th>
<th>Organism</th>
<th>Sampling frequency</th>
<th>Number of components</th>
<th>Analytical methods and measurement error</th>
<th>References</th>
</tr>
</thead>
<tbody>
<tr>
<td>fast sampling</td>
<td>microorganisms</td>
<td>sub second</td>
<td>5-15 metabolites</td>
<td>enzymatic analysis; HPLC; variation duplicates &lt; 3%; variation between different extractions &lt; 10%;</td>
<td>[40, 20]</td>
</tr>
<tr>
<td>from bioreactor (stopped-flow technique)</td>
<td>(E. coli, yeast)</td>
<td></td>
<td></td>
<td></td>
<td>[210]</td>
</tr>
</tbody>
</table>

18
## Reverse engineering of metabolic networks, a critical assessment

<table>
<thead>
<tr>
<th>Fast sampling</th>
<th>Micro-organisms</th>
<th>Sub second</th>
<th>15-30 metabolites</th>
<th>Enzymatic analysis</th>
</tr>
</thead>
<tbody>
<tr>
<td>from E.coli</td>
<td></td>
<td></td>
<td>HPLC;</td>
<td>[169]</td>
</tr>
<tr>
<td>Bioreactor</td>
<td></td>
<td></td>
<td>LC-MS-MS;</td>
<td>[18]</td>
</tr>
<tr>
<td>(Automated</td>
<td></td>
<td></td>
<td>RSD</td>
<td>[210]</td>
</tr>
<tr>
<td>Sampling</td>
<td></td>
<td></td>
<td>= 5 – 10%</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Fast sampling</th>
<th>Micro-seconds</th>
<th>5-20 metabolites</th>
<th>Enzymatic analysis;</th>
</tr>
</thead>
<tbody>
<tr>
<td>from yeast</td>
<td></td>
<td></td>
<td>HPLC;</td>
</tr>
<tr>
<td>Bioreactor</td>
<td></td>
<td></td>
<td>LC-MS-MS;</td>
</tr>
<tr>
<td>(Bioscope)</td>
<td></td>
<td></td>
<td>UV-VIS;</td>
</tr>
<tr>
<td>(Bioscope)</td>
<td></td>
<td></td>
<td>GC-FID;</td>
</tr>
<tr>
<td>(Bioscope)</td>
<td></td>
<td></td>
<td>LC-ESI-MS-MS;</td>
</tr>
<tr>
<td>(Bioscope)</td>
<td></td>
<td></td>
<td>(IDMS);</td>
</tr>
<tr>
<td>(Bioscope)</td>
<td></td>
<td></td>
<td>RSD</td>
</tr>
<tr>
<td>(Bioscope)</td>
<td></td>
<td></td>
<td>= 5 – 25%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Fast sampling</th>
<th>Micro-seconds</th>
<th>5-20 metabolites</th>
<th>Enzymatic analysis;</th>
</tr>
</thead>
<tbody>
<tr>
<td>from yeast</td>
<td></td>
<td></td>
<td>HPLC;</td>
</tr>
<tr>
<td>Bioreactor</td>
<td></td>
<td></td>
<td>Bio-luminescence;</td>
</tr>
<tr>
<td>(Stimulus</td>
<td></td>
<td></td>
<td>Variation</td>
</tr>
<tr>
<td>Response</td>
<td></td>
<td></td>
<td>Between triplicates</td>
</tr>
<tr>
<td>Approach</td>
<td></td>
<td></td>
<td>&lt; 5%;</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Variation</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Between different</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Extractions</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>&lt; 10%;</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Error bio-luminescence</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>≥ 20%;</td>
</tr>
</tbody>
</table>
Reverse engineering of metabolic networks, a critical assessment

| slow sampling organisms from bioreactor | micro- hours 10-15 metabolites | error HPLC $\leq 6\%$ | LC-ESI-MS-MS (IDMS); $= 10 - 30\%$ |
| slow sampling organisms from bioreactor | micro- hours tens or hundreds metabolites | GC-MS; IP-LC-MS; OS-GC-MS; $= 5 - 20\%$ |
| tissue samples plants metabolites | minutes/ hours 15-20 GC-MS; LC-MS; NMR; RSD $= 10 - 20\%$ |
| blood samples (males) leptin: hormones 20 minutes | human 10 minutes; 1-8 inter assay variation $< 20\%$; intra-assay variation $< 5\%$; RSD $= 5 - 25\%$ |
| blood samples (liver) | dog, rat 45 minutes 4 lipids GC-MS; RSD $= 5 - 10\%$ |
| blood samples (females) hormones minutes/ hours 5-10 | human inter assay variation $< 15\%$; intra-assay variation immunoassay’s; $97, 96$ |
Reverse engineering of metabolic networks, a critical assessment

| blood samples | human minutes/ hours | hundreds (lipids) | LC-MS; [11] RSD | < 7%; 5 – 25% |

Abbreviations: HPLC, High Performance Liquid Chromatography; LC-MS, Liquid chromatography-mass spectrometry; UV-VIS, ultraviolet-visible spectrophotometry; GC-FID, Gas Chromatography-Flame Ionization Detector; LC-ESI-MS-MS, Liquid chromatography electro spray ionization tandem mass spectrometry; IDMS, Isotope Dilution Mass Spectrometry; GC-MS, Gas chromatography-mass spectrometry; IP-LC-MS, ion-pair liquid chromatography-mass spectrometry; OS-GC-MS, Oximation silylation-gas chromatography-mass spectrometry; NMR, Nuclear Magnetic Resonance; RSD, relative standard deviation.
Table 2.2: Properties of the network inference methods, described in this chapter.

<table>
<thead>
<tr>
<th>Property</th>
<th>Partial Pearson correlations</th>
<th>Time-lagged correlations</th>
<th>Penalized Jacobian method</th>
<th>Zero slopes method</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data</td>
<td>steady state</td>
<td>dynamic</td>
<td>dynamic</td>
<td>dynamic</td>
</tr>
<tr>
<td>Temporal information</td>
<td>none</td>
<td>whole time</td>
<td>whole time profile</td>
<td>only first</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>measurements</td>
</tr>
<tr>
<td>Topology</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
</tr>
<tr>
<td>Directionality</td>
<td>no</td>
<td>no</td>
<td>yes</td>
<td>yes</td>
</tr>
<tr>
<td>Interaction strength</td>
<td>no</td>
<td>no</td>
<td>yes</td>
<td>no</td>
</tr>
<tr>
<td>Perturbations</td>
<td>biological variation</td>
<td>perturbation of the</td>
<td>only small perturbations</td>
<td>one metabolites</td>
</tr>
<tr>
<td></td>
<td></td>
<td>carbon source at regular</td>
<td></td>
<td>at a time</td>
</tr>
<tr>
<td></td>
<td></td>
<td>time-intervals</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sampling frequency</td>
<td>not important</td>
<td>fast or intermediate</td>
<td>very fast</td>
<td>fastest</td>
</tr>
<tr>
<td>Number of metabolites to be</td>
<td>all</td>
<td>all</td>
<td>all</td>
<td>all</td>
</tr>
<tr>
<td>measured</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Number of samples</td>
<td>very high</td>
<td>intermediate</td>
<td>high</td>
<td>high</td>
</tr>
<tr>
<td>Section</td>
<td>2.2.1.</td>
<td>2.2.2.</td>
<td>2.2.3.</td>
<td>2.2.4.</td>
</tr>
</tbody>
</table>

\textsuperscript{a}The interaction strength is a measure of how strong two metabolites influence each other. For a detailed definition, see methods section.
2.2 Methods

A metabolic network is a graph where the metabolites are represented by nodes and the (direct) interactions between them by edges. The directionality of the network is indicated by arrows, where an arrow from one metabolite to another means that changes in the concentrations of the first metabolite have a direct influence on the other.

The time behavior of most metabolic networks is non-linear and can be written as follows [88]:

$$\frac{dS_i(t)}{dt} = F_i(S_1(t), \cdots, S_n(t))$$ (2.1)

for $i = 1, \cdots, n$ where $n$ is the number of metabolites in the network, $d(.)/dt$ the time derivative, $S_1(t), \cdots, S_n(t)$ the concentration profiles of the metabolites and $F_i(.)$ a function that describes the effects of $S_1(t), \cdots, S_n(t)$ on $S_i(t)$.

The estimation of the derivative of the concentration with respect to time can be linked to the calculation of the Jacobian matrix [37, 192]. The entries of the Jacobian matrix ($i, j \in \{1, \cdots, n\}$) can be expressed as:

$$J_{ji} = \frac{\partial F_j}{\partial S_i}$$ (2.2)

If metabolite $i$ influences metabolite $j$, then $J_{ji} \neq 0$ and if $J_{ji} = 0$, then $i$ does not affect $j$. For each metabolite pair $i$ and $j$, the absolute maximum of the entries $J_{ij}$ and $J_{ji}$ is called the interaction strength between metabolite $i$ and metabolite $j$ [27]. For an unknown network, the connections between the metabolites and the direction of the edges in the network can be deduced by calculating the Jacobian.

The number of parameters (Jacobian matrix entries) increases quadratically with the number of metabolites in the network. Unreasonable computational time is the result of wanting to calculate the Jacobian matrix for large(r), non-linear systems. As a consequence, it is difficult to find the interactions between the metabolites in the network.
2.2.1 Similarity measures

Çakır and coworkers [27] did an analysis of metabolic network inference from in silico metabolomic datasets based on statistical similarity measures. They analyzed different types of data based on variability around steady state. This study focuses on two types of variability: enzymatic and intrinsic variability. Enzymatic variability is caused by small variations of enzyme concentrations or reaction constants between replicate experiments. With intrinsic variability we mean fluctuations within cellular processes. These fluctuations are due to changes in the intracellular milieu [27]. Çakır and coworkers [27] showed that n-th order partial Pearson correlation (PPCn) is the best similarity measure to use for metabolic network inference. The method is illustrated in Figure 2.1. Available steady state data of the metabolite concentrations are collected. Pearson correlation coefficients are calculated for all pairwise plots of metabolite concentrations. A high correlation coefficient can indicate both a direct or an indirect interaction between two metabolites. To discriminate between these two types of interaction, a conditioned similarity measure is used: the n-th order partial Pearson correlation. This measure describes the correlation between two metabolites, conditioned on the remainder of the metabolites [170]. The details about the calculation of partial Pearson correlations can be found in Appendix A.

2.2.2 Time-lagged correlations

Correlations can be used to infer biological networks from time series data. The influence of one species on another is often observed after a time delay. As a consequence, it can happen that two time series have a low correlation coefficient while there is a strong correlation between them if a time lag is allowed (see Figure 2.2).

Arkin and Ross (1995) [5] proposed a method to infer biological networks from time series by computing a time-lagged correlation matrix. Details about the calculations are explained in Appendix B. Figure 2.3 gives a schematic overview of the method. Available time series resulting from perturbation of the carbon source at regular time intervals are collected. Time-lagged correlation coefficients are calculated for all
Reverse engineering of metabolic networks, a critical assessment

Figure 2.1: Illustration of the reverse engineering method based on partial Pearson correlation. Adapted from Çakır and coworkers [27].

Pairwise plots of metabolite concentrations (see Appendix B). A threshold has to be specified to distinguish between metabolites that have a direct interaction and those that have not. In this study, a threshold was determined by looking up the sparsity of metabolic networks of the same size as the ones studied in the JWS online models database [146]. For networks of 13-18 metabolites, the ratio of the number of real interactions in the network to the number of possible interactions is around 0.25. Because of this, the third quartile of the time-lagged correlation matrix was chosen as threshold. For studying networks of another size, another percentile has to be chosen as cut-off value in the way described above.

2.2.3 The penalized Jacobian method

If we focus on small perturbations from steady state, equation (2.1) can be simplified by a linear approximation [37]:

\[
\frac{dS}{dt} \approx J \cdot (S - S_0)
\]  

(2.3)
where $S_0$ is the matrix of steady state concentrations. The derivatives $\frac{dS}{dt}$ were numerically calculated using a fourth order approximation (see Appendix C). This makes the system in (2.1) linear. The Jacobian matrix can be estimated from the system above with an algorithm based on least squares estimates. Because metabolic networks are sparse, a penalty was used to get a sparse solution of the system. Details about the algorithm are given in Appendix C. In practice, the obtained Jacobian has no zero entries, but entries close to zero. A cut-off value $\delta$ needs to be specified. If an entry in the Jacobian is smaller than $\delta$, than it is treated as a zero.

Figure 2.4 gives a schematic presentation of the Jacobian method. Available time series resulting from small perturbations of one or more metabolites from steady state are collected. From this data, the Jacobian matrix is calculated. For each pair of metabolites $i$ and $j$, the element in the $i$-th column, $j$-th row describes how changes in metabolite $i$ affect changes
in metabolite j. If this element is non-zero, there is an edge from i to j in the network. In this way, a directed graph can be derived from the Jacobian.

### 2.2.4 The zero slopes method

A possible solution to find the connectivity and directionality of metabolic networks is to use zero slope information [213, 37]. Figure 2.5 gives an overview of the method. Available time series resulting from increasing the metabolites from steady state one at a time were collected. These concentration profiles are analyzed. In the curves of the concentrations, different kinds of profiles can be observed. The curve of the increased metabolite A decreases monotonically. For the concentration profiles of the remaining metabolites, there are three possibilities. First, the curve of a metabolite B can be a constant graph, which means that A has no influence on B. Second, an increasing curve of metabolite B with non-zero initial slope can be observed, which means that A has a direct effect on B. In this case an arrow is drawn from A to B. Finally, the curve of metabolite B can have an initial zero slope, indicating that A has an indirect effect on B (A acts on B through a third metabolite). With the information derived from the concentration profiles, the vertex-edge incidence matrix N can be constructed (see Appendix D). This is a matrix of zeros and ones, where a 1 in the i-th column, j-th row indicates that
metabolite i has a direct effect on metabolite j and a 0 means that this is not the case. A directed graph is constructed from the vertex-incidence matrix.

### 2.2.5 Method performance

There are different ways to measure the quality of a network inference method. In this study, the geometric mean score (g-score) is used, which is the geometric mean of the sensitivity (the true positive rate TPR) and the specificity (the true negative rate TNR) [27]:

\[
g\text{- score} = \sqrt{\text{sensitivity} \times \text{specificity}} = \sqrt{TPR \times TNR}
\]

The g-score is always a number between 0 and 1, where a g-score of 1 corresponds with perfect inference.
Figure 2.5: Overview of the zero slopes method.

The information in the appendix of Çakır and coworkers [27] is used to calculate the g-scores for the PPCn method.

2.2.6 Effect of noise on network inference

The influence of noise in the data on the performance of the inference methods was examined. This was done by adding noise drawn from a random normal distribution, with zero mean, to the measurements by putting the standard deviation of the normal distribution to 3% of the noiseless data. A hundred datasets with this noise level were simulated. Before calculating derivatives, the data were smoothed with cubic splines [64]. The reverse engineering methods were applied on the hundred
datasets and the mean of the g-scores was calculated.

2.2.7 Simulations

Matlab’s [127] ordinary differential equation solver *ode15s* (The Mathworks, MATLAB Version 7.5.0., 2007, Microsoft Windows XP Version 5.1) was used for simulating concentration profiles and determining steady state concentrations. Kinetic parameters and differential equations were taken from Teusink and coworkers [201] for *S. cerevisiae* and from Chasagnole and coworkers [30] for *E. coli*. Simulations for the method based on similarity measures are performed like described in Çakir and coworkers [27].

For the time-lagged correlation method, simulations mimicking the experiments described by Arkin and coworkers [6] are performed. An initial glucose concentration that is 1-3% increased from steady state was chosen. All other metabolites were initially at steady state. After a time interval that is long enough to bring the system back at steady state, a new glucose-perturbation of 1-3% was performed. In this way hundred time-points were simulated.

The approximations for the Jacobian method are only valid for small perturbations from steady state (see paragraph 2.2.3.). Furthermore, a higher number of time courses increases the performance of the method [184]. Therefore as many experiments as metabolites in the network were performed as follows. One metabolite had an initial concentration that was 2% higher than steady state while all the other metabolites were initially at steady state. Again, each time series consists of one hundred time points.

Time series simulated for studying the zero slopes method are similar as those for the Jacobian method, but with a perturbation of 10% from steady state. For time profiles that are further away from equilibrium it is easier to distinguish between a zero and a non-zero slope.

For each method, experiments were performed for different time intervals between the measurements to study the influence of the sampling frequency.
2.3 Results and discussion

2.3.1 Results

The four methods discussed above were evaluated for two metabolic networks: the glycolytic pathway of *S. cerevisiae* and the central carbon metabolism of *E. coli*. The glycolytic pathway of *S. cerevisiae* is a network of 13 metabolites and 18 reactions [201]. There are 78 possible pairwise interactions, of which 21 are in the real network [27]. The maximum number of possible arrows in the network is 156, while 41 are occurring in the real network [201]. The central carbon metabolism of *E. coli* is a network that consists of 18 metabolites and 30 reactions [30]. This network has 153 possible pairwise interactions, whereas the real network has 37 [27]. 306 arrows between the metabolites are possible, while the real number is 72 [30]. The results for both networks are shown in Table 2.3, additional properties derived from the results are summarized in Table 2.4.

From Table 2.3 it can be deduced that a complete recovery of the network, given the perturbations and sampling frequencies used to generate the simulated data, is impossible (g-scores, TNR and most TPR are smaller than 1). This means that edges in the network are missing but also that the inferred network contains edges that are not found in the real one. The partial Pearson correlations, the time-lagged correlations and Jacobian method could prune the networks reasonably well (high TNR). The zero slopes method finds most edges in the network (high TPR), but is in some cases (e.g. *S. cerevisiae*) less capable to remove indirect interactions (lower TNR). An indirect interaction between two metabolites i and j means that two or more intermediate reactions are needed to form j from i.

The time-lagged correlations and in some cases also the Jacobian method can only infer a small fraction of the real interactions in the network (low TPR).

Although the central carbon metabolism of *E. coli* consists of five more metabolites and twelve more reactions than the glycolytic pathway of *S. cerevisiae*, the g-scores are in general higher for the former.

Performance of all methods decreases dramatically when noise is added

31
to the data. The g-scores for the Jacobian method are especially low, meaning that noise has the largest effect on this method. Furthermore, Table 2.3 shows that g-scores for the Jacobian and zero slopes method increase with the sampling frequency, indicating that these methods perform better if sampling is done faster. Results for *E. coli* and *S. cerevisiae* are comparable.
<table>
<thead>
<tr>
<th>method</th>
<th>S. cerevisiae</th>
<th>E. coli</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>samples/∆t</td>
<td>noiseless</td>
</tr>
<tr>
<td></td>
<td></td>
<td>g-score</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(TPR,TNR)</td>
</tr>
<tr>
<td>partial</td>
<td>enzymatic</td>
<td>0.75</td>
</tr>
<tr>
<td>Pearson</td>
<td>variability</td>
<td>(0.69,0.81)</td>
</tr>
<tr>
<td></td>
<td>intrinsic</td>
<td>0.82</td>
</tr>
<tr>
<td></td>
<td>variability</td>
<td>(0.77,0.87)</td>
</tr>
<tr>
<td>time-lagged</td>
<td>∆t</td>
<td>0.63</td>
</tr>
<tr>
<td></td>
<td>= 0.2 min</td>
<td>(0.48,0.82)</td>
</tr>
<tr>
<td></td>
<td>correlations</td>
<td>Δt</td>
</tr>
<tr>
<td></td>
<td>= 0.1 min</td>
<td>(0.48,0.84)</td>
</tr>
</tbody>
</table>

b For details about the four approaches see Methods section.

∆t is the time between two measurements.

do Abbreviations: g-score = geometric mean score; TPR = true positive rate; TNR = true negative rate. For an explanation see Methods section.

e See Methods.

f Results for noiseless data based on the appendix of Çakir and coworkers [27].

g Results for noiseless data based on the appendix of Çakir and coworkers [27].

h See Methods.

i Results for noiseless data based on the appendix of Çakir and coworkers [27].

j Results for noiseless data based on the appendix of Çakir and coworkers [27].

k The same time unit as Chassagnole and coworkers [30] was used.

l The same time unit as Teusink and coworkers [201] was used.
<table>
<thead>
<tr>
<th></th>
<th>$\Delta t$</th>
<th>$\Delta t$</th>
<th>$\Delta t$</th>
<th>$\Delta t$</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>penalized</strong></td>
<td>0.33</td>
<td>0.27</td>
<td>0.55</td>
<td>0.04</td>
</tr>
<tr>
<td><strong>Jacobian</strong></td>
<td>$10^{-3}$</td>
<td>$10^{-2}$</td>
<td>$10^{-3}$</td>
<td>$10^{-3}$</td>
</tr>
<tr>
<td><strong>method</strong></td>
<td>0.46-0.71</td>
<td>0.73-0.75</td>
<td>0.92-0.75</td>
<td>0.85-0.97</td>
</tr>
<tr>
<td><strong>zero slopes</strong></td>
<td>0.65</td>
<td>0.48</td>
<td>0.88</td>
<td>0.61</td>
</tr>
<tr>
<td><strong>method</strong></td>
<td>$10^{-3}$</td>
<td>$10^{-2}$</td>
<td>$10^{-3}$</td>
<td>$10^{-3}$</td>
</tr>
<tr>
<td></td>
<td>0.95-0.44</td>
<td>0.83-0.93</td>
<td>0.85-0.97</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.77</td>
<td>0.91</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>1.05</td>
<td>0.62</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Reverse engineering of metabolic networks, a critical assessment

Table 2.4: Properties of the network inference methods derived from the results.

<table>
<thead>
<tr>
<th>Property</th>
<th>Partial Pearson correlations</th>
<th>Time-lagged correlations</th>
<th>Penalized Jacobian method</th>
<th>Zero slopes method</th>
</tr>
</thead>
<tbody>
<tr>
<td>Distinguish between direct and indirect interactions</td>
<td>reasonable</td>
<td>reasonable</td>
<td>for large sampling frequencies</td>
<td>for large sampling frequencies</td>
</tr>
<tr>
<td>Influence of noise</td>
<td>large</td>
<td>large</td>
<td>very large</td>
<td>large</td>
</tr>
</tbody>
</table>

2.3.2 Discussion

Applying the network inference approaches described in this chapter, two phenomena can be observed. On the one hand, often edges are missing in the inferred network, but on the other hand also spurious connections are found. The missing edges usually correspond with the weakest interactions in the network, as already discussed by Çakir [27]. Interactions are called weak if their interaction strength is lower than 1 (for a definition of the interaction strength, see Methods section)[27]. The spurious edges are often the result of the incapability of methods to discriminate between direct and indirect interactions. Very fast indirect interactions can lead to edges in the inferred network that do not exist in the real one.

For the penalized Jacobian method, the true positive rates for *E. coli* are in general higher than for the smaller network of *S. cerevisiae*. A possible explanation for this could be that the *E. coli* network is sparser (ratio between the number of occurring connections and the number of possible connections is smaller) than that of *S. cerevisiae*. A method that pushes the system into a sparse solution gives less false negatives for a sparser network.

The true negative rates for the zero slopes method are generally higher for *E. coli* than for *S. cerevisiae*. The fastest interactions in the network of *E. coli* have an interaction strength in the order of $10^4$, which is much smaller than that for *S. cerevisiae* that is of order $10^6$. Successions of
these very fast interactions result in many false positives when estimating the network in *S. cerevisiae*.

The performance of each of the network inference methods is, although in different amounts, enormously affected by noise. Noise levels of 3% which are lower than those of real experiments (in the order of 5-25%, see Table 2.1), already lead to an enormous decrease of the performance of the methods. This can only be solved by new measurement techniques with lower sampling and measurement errors.

For recovering fast reactions, all time series based methods, except time-lagged correlations, require fast sampling which can be illustrated by the following hypothetical experiment. A dataset with a very large sampling frequency is simulated for *S. cerevisiae* (e.g. $\Delta t = 10^{-9}$ minutes). For this experiment the complete network can be inferred with the Jacobian and zero slopes method. However, taking smaller sampling frequencies led to false positives or false negatives for the zero slopes method. For the Jacobian method g-scores below 1 were found when the sampling frequency was lower than $10^{-6}$ minutes, which means that the zero slopes method asks for the fastest measurement frequencies. This was also noticed by Crampin and coworkers [37]. Time-lagged correlations did not show better performance if faster sample frequencies were simulated. If the reactions are too fast, required sampling rates are not possible with current laboratory techniques. Also Crampin and coworkers [37] mentioned that sampling must be sufficiently fast, but they did not elucidate yet how fast.

If the network consists of fast and slow reactions, then a fast sampling scheme has to be used for a long time. Fast sampling is needed because of reasons explained above. Sampling a long time is needed to estimate the connectivity due to slow reactions. This means that for a network consisting of a large range of different reaction constants, very many samples have to be taken, which is also discussed by Delgado-Eckert [41]. This has severe repercussions for designing experiments (see below).

Another disadvantage of all the methods described in this chapter is that all metabolites in the network have to be measured. In practice, it is often not possible to do this [37]. The zero slopes method can only be used if the metabolites are increased one at a time, which creates a
large experimental burden and is therefore not very practical.

Recent work of Srividhya and coworkers [184] showed similar problems for another network inference method that selects chemical reactions that best fit the data. This method was less affected by noise than the four methods in this work, but other problems occurred. One of them was that more than one set of chemical reactions can fit the data equally well. Furthermore, the number of reactions increases exponentially with the number of metabolites, which makes this method computationally intractable for larger networks.

From all of the points mentioned above it can be concluded that with the analytical technology and the network inference methods of today it is not possible to infer a whole large network without a substantial amount of errors. On the experimental part, measurement methods have to improve considerably in terms of noise. To design a time-resolved experiment for reverse engineering it is very important to have some prior information regarding the sizes of the reaction constants to expect. This should focus the design. In terms of reverse engineering methods, fortunately, directions to other solutions exist, since already a priori information about metabolic networks exists in metabolic databases [50]. This knowledge can then be used in a specific situation to infer a metabolic network, e.g. by using grey models [225]. This route will be explored in our group in subsequent work.

2.4 Conclusion

The described inference methods extract information on the interaction network from experimental metabolite concentration data. None of the approaches presented in this chapter offers network inference without errors, although a thought experiment with much higher sampling frequency than those of recent measurement techniques pointed out that the zero slopes and Jacobian method could in principle infer the whole network. For time-lagged correlations, this was not the case. There may be a huge difference between the smallest and the largest time constant in the network. In general, this difference becomes larger if the network contains more metabolites. As a consequence, sampling
fast and for a long period is needed to observe both the fast and the slow interactions. This study showed that sometimes the fastest interactions can not be inferred from the data obtained with recent analytical techniques.

Noise in the data is an important factor influencing the performance of any of the proposed techniques. Current analytical techniques have high measurement errors (up to 25%, see Table 2.1). If the measurement noise could be reduced, this would contribute to a better network inference.

Time series may contain not enough information to apply the proposed inference methods. This is the case when the slowest interactions did not take place during the observed time interval or when some interactions are too fast to be measured with current laboratory techniques.

In summary, it can be concluded that if full inference of a large metabolic network is the goal then the requirements for the sampling frequency are not consistent with contemporary practice. A similar result for the inference of gene networks can be found in recent literature [63]. However, it is not needed to estimate the whole network from the data because there exist already a lot of biological information in databases [50]. Information on the order of the reaction rates, known parts of the network and modules in the network can greatly improve network inference. Moreover, this information can help to set up experiments that give more informative data (e.g. sampling long enough to capture the slowest interactions). Integration of the bottom-up approach of building networks from knowledge deposited in databases and the top-down approach (reverse engineering) could be an option for further research. By incorporating biological knowledge, maximal information can be extracted about the network, given the current data.
Appendix A: Calculation of n-th order partial Pearson correlations

The Pearson correlation coefficient between two metabolites \( i \) and \( j \) is defined by [37]:

\[
\rho_{ij} = \frac{\sum_{k=1}^{m} \left( \frac{S_i(k) - \bar{S}_i}{\sigma_i} \right) \left( \frac{S_j(k) - \bar{S}_j}{\sigma_j} \right)}{m}
\]  

(2.4)

where \( m \) is the number of samples; \( S_i \) and \( S_j \) are the concentrations of metabolites \( i \) and \( j \); \( \bar{S}_i \) and \( \bar{S}_j \) are the mean values of \( S_i \) and \( S_j \); \( \sigma_i \) and \( \sigma_j \) are the sample standard deviations of \( S_i \) and \( S_j \). The Pearson correlation coefficients are the entries of the Pearson correlation matrix \( P = (\rho_{ij}) \). The entries of the n-th order partial correlation matrix \( \Pi = (\pi_{ij}) \) describe the correlation between any two metabolites \( i \) and \( j \) conditioned on all remaining metabolites. The matrix \( \Pi \) can be calculated from the matrix \( P \) with the following two formulas [170]:

\[
\Omega = P^{-1} = (\omega_{ij})
\]  

(2.5)

\[
\Pi_{ij} = \frac{-\omega_{ij}}{\sqrt{\omega_{ii} \cdot \omega_{jj}}}
\]  

(2.6)

A significance measure for the similarity scores was calculated by performing a permutation test like described by Çakir and coworkers [27]. The final network consists of all edges with significant similarity scores.

Appendix B: Calculation of the time-lagged correlation matrix

For a time series of length \( m \), the cross-covariance is calculated for all \( \frac{1-m}{2} \leq \tau \leq \frac{m-1}{2} \) with the formula [2]:

\[
\phi_{ij}(\tau) = \sum_{k=1}^{m-\tau} (S_i(t_k) - \bar{S}_i) \cdot (S_j(t_{k+\tau}) - \bar{S}_j)
\]
Reverse engineering of metabolic networks, a critical assessment

if $\tau \geq 0$

$$\phi_{ij} (\tau) = \sum_{k=1-\tau}^{m} (S_i (t_k) - \bar{S}_i) \cdot (S_j (t_{k+\tau}) - \bar{S}_j)$$

if $\tau < 0$

where $i, j \in \{1, \ldots, n\}$ with $n$ the number of metabolites in the network, $\Delta t$ = the time between two measurements and $t_{k+\tau} = t_k + \tau \Delta t$.

The time-lagged correlation coefficients for all $\frac{1-m}{2} \leq \tau \leq \frac{m-1}{2}$ can then be calculated with the following formula [2]:

$$C_{ij} (\tau) = \frac{\phi_{ij} (\tau)}{\sqrt{\phi_{ii} \cdot \phi_{jj}}}$$

where $\phi_{ii}$ and $\phi_{jj}$ are the standard deviations of the $m - \tau$ data points included in the formula for $\phi_{ij} (\tau)$, calculated for $S_i$ and $S_j$ respectively. The entries of the time-lagged correlation matrix are determined by taking the maximal absolute time-lagged correlation between metabolites $i$ and $j$ [5]:

$$c_{ij} = max_{\frac{1-m}{2} \leq \tau \leq \frac{m-1}{2}} |C_{ij} (\tau)|$$

where $i, j \in \{1, \ldots, n\}$.

**Appendix C: Algorithm to calculate the Jacobian matrix**

If we focus on small perturbations from steady state, equation (2.1) can be simplified by a linear approximation [37]:

$$\frac{dS}{dt} \approx J \cdot (S - S_0)$$

where $S_0$ is the matrix of steady state concentrations. The derivatives $\frac{dS}{dt}$ were numerically calculated using a fourth order approximation:

$$\frac{dS (t_n)}{dt} \approx -S (t_n + 2\Delta t) \frac{12 \cdot \Delta t}{12 \cdot \Delta t} + 8S (t_n + \Delta t) \frac{12 \cdot \Delta t}{12 \cdot \Delta t} - 8S (t_n - \Delta t) \frac{12 \cdot \Delta t}{12 \cdot \Delta t} + S (t_n - 2\Delta t) \frac{12 \cdot \Delta t}{12 \cdot \Delta t}$$
where \( S(t_n) \) is the concentration at time \( t = t_n \) and \( \Delta t \) is the time between two measurements. This makes the system in (2.1) linear.

**Least squares.**

The Jacobian matrix can be estimated using least squares estimates, which means minimizing the residual sum of squares

\[
\|XJ^T - Y\|_2^2
\]

where \( J^T \) is the transpose of the Jacobian matrix, \( X = S - S_0 \) is a matrix containing the concentrations minus the steady state value for the different time points and \( Y = \frac{dS}{dt} \) is a matrix that contains the approximated derivatives of the concentrations.

Least squares estimation can be done by using the following formula:

\[
J^T = (X^T X)^{-1} X^T Y
\]

The matrix \( X^T X \) may be singular, so that it has no inverse. To overcome this problem, ridge regression is used. A small positive number \( \gamma \) is first added to the diagonal of \( X^T X \), before the matrix is inverted:

\[
J^T = (X^T X + \gamma I)^{-1} X^T Y
\]

**L1 regularization.**

The methods mentioned above create a solution with most entries in the Jacobian non-zero. Because metabolic networks are sparse, a penalty was used to push the system \( Y = XJ^T \) into a sparse solution. This can be done by using an L1-penalty, which means minimizing [173]:

\[
\|XJ^T - Y\|_2^2 + \lambda \|J^T\|_1 \tag{2.9}
\]

This problem is also known as the LASSO (Least Absolute Selection and Shrinkage Operator) minimization problem.
**L0 + L1 regularization.**

A disadvantage of L1 regularization is that there are often still too many edges in the solution. This can be overcome by using an L1 penalty combined with an L0 penalty [47]. There are a lot of approaches proposed for this type of minimization problems. An iterated approach proposed by Schmidt [173] is used. It starts with an initial estimate J of the Jacobian, for instance obtained by least squares or ridge regression. The vec operator was applied to this initial estimate, which means that the columns are placed one underneath the other, so that a vector 

$$j = vec(J^T) = [j_1, ..., j_n]^T$$

is obtained, where n is the number of entries in the Jacobian. From this initial estimate, a better estimate is calculated by using the formula

$$j_{new} = (X^T X + \Gamma^T \Gamma)^{-1} X^T y$$

(2.10)

where \( y = vec(Y) \) and

$$\Gamma^T \Gamma = \lambda^2 \begin{bmatrix}
\frac{1}{\epsilon + |j_1| + \kappa j_1^2} & 0 & \cdots & 0 \\
0 & \ddots & 0 & \vdots \\
\vdots & 0 & \ddots & 0 \\
0 & \cdots & 0 & \frac{1}{\epsilon + |j_n| + \kappa j_n^2}
\end{bmatrix}$$

(2.11)

where \( \epsilon \) is a very small number \((10^{-6} - 10^{-4})\) and \( \lambda \) and \( \kappa \) penalty parameters [47].

The last step is repeated until convergence. This can be performed by using a stop criterion. A possible stop criterion could be the following. If the sum of squares of \( \frac{J_{old} - J_{new}}{J_{old}} \) is lower than 0.0001, stop the program, else perform a next iteration step, where \( J_{old} \) and \( J_{new} \) are the Jacobians obtained by the previous and the current iteration step respectively.

Because we know *a priori* that a metabolite has an effect on itself, the penalty was not applied to the diagonal elements of the Jacobian.
Appendix D: The vertex-edge incidence matrix

Example

Suppose the hypothetical network depicted in Figure 2.6 has to be modeled.

The mass balances for each metabolite are:

\[
\frac{dS_1}{dt} = k_1 - k_2 S_1 - k_4 S_1 + k_3 S_3
\]

\[
\frac{dS_2}{dt} = -k_6 S_2 - k_7 S_2 + k_5 S_4
\]

\[
\frac{dS_3}{dt} = -k_3 S_3 + k_2 S_1 + k_6 S_2
\]

\[
\frac{dS_4}{dt} = k_4 S_1 - k_5 S_4
\]

We took as an example \( k_1 = k_3 = k_5 = k_7 = 1 \) and \( k_2 = k_4 = k_6 = 2 \).

Figure 2.6: A hypothetical network, consisting of four metabolites.

The metabolites were increased from steady state one at a time to simulate four time series. In the curves of the concentrations, different
kinds of profiles can be observed (see Figure 2.7). The curve of the increased metabolite decreases monotonically. An increasing curve with non-zero initial slope means that there is a direct interaction between the increased metabolite and the metabolite which concentration profile is presented in the curve. A zero slope means that a change in the increased metabolite needs some time to manifest itself in the metabolite presented by the curve. This is because two or more intermediate reactions are needed to form this metabolite. This is called an indirect interaction. Finally it can also happen that the concentration profile is constant (not shown in the figure). This means that the increased metabolite has no influence on the metabolite in the curve.

Figure 2.7: Concentration profiles for the system in Figure 2.6 when the initial concentration of $S_2$ is higher than steady state and all the other initial concentrations are at steady state. Zero slopes are indicated with black circles.
Reverse engineering of metabolic networks, a critical assessment

General

With the information above, the vertex-edge incidence matrix $N = \{N_{ij}\}(i, j \in \{1, \cdots, n\})$ was defined as follows:

- If $S_i$ is increased, the term in $S_i$ in the differential equation for $\frac{dS_i}{dt}$ is negative and $N_{ii} = -1$.

- If increasing $S_i$ does not influence $S_j$, there is no term in $S_i$ in the differential equation for $\frac{dS_j}{dt}$ and $N_{ji} = 0$.

- If there is a zero slope on the graph of $S_j$ when $S_i$ is increased, there is no term in $S_i$ in the differential equation for $\frac{dS_j}{dt}$ and $N_{ji} = 0$.

- If the graph of $S_j$ increases and after reaching a maximum decreases to steady state when $S_i$ is increased, the term in $S_i$ in the differential equation for $\frac{dS_j}{dt}$ is positive and $N_{ji} = 1$.

In practice, initial slopes will not be zero but very close to zero. In that case a cut-off value $\eta$ has to be specified. Initial slopes smaller than $\eta$ are treated as a zero.

Acknowledgments

This project was financed by the Netherlands Metabolomics Centre (NMC) which is part of the Netherlands Genomics Initiative / Netherlands Organization for Scientific Research. Tunahan Çakır is gratefully acknowledged for providing us with Matlab code on partial Pearson correlations and powerpoint slides that we could use for making parts of the figures in this chapter. We thank Daniël J. Vis (University Medical Centre Utrecht and University of Amsterdam) for his comments on the manuscript.
Reverse engineering of metabolic networks, a critical assessment
Chapter 3

Global test for metabolic pathway differences between conditions

In many metabolomics applications there is a need to compare metabolite levels between different conditions, e.g., case versus control. There exist many statistical methods to perform such comparisons but only few of these explicitly take into account the fact that metabolites are connected in pathways or modules. Such *a priori* information on pathway structure can alleviate problems in, e.g., testing on individual metabolite level. In gene-expression analysis, Goeman’s global test is used to this extent to determine whether a group of genes has a different expression pattern under changed conditions. We examined if this test can be generalized to metabolomics data. The goal is to determine if the behavior of a group of metabolites, belonging to the same pathway, is significantly related to a particular outcome of interest, e.g., case/control or envi-

---

The results show that the global test can indeed be used in such situations. This is illustrated with extensive intracellular metabolomics data from *E. coli* and *S. cerevisiae* under different environmental conditions.

### 3.1 Introduction

Many current problems in metabolomics can be summarized as finding differences between conditions. The prototypical metabolomics biomarker study is an example: diseased versus control individuals are subjected to urine or serum metabolomics measurements and subsequently statistical methods are used to find the differences. This is mostly done using multivariate data analysis tools such as PLS-DA (Partial Least Squares Discriminant Analysis) [138, 226], but also univariate tools are used [69]. Both tools have drawbacks, e.g., in univariate methods the multiple testing problem is present and in multivariate analysis model interpretation can be difficult. Shortcuts have been proposed, such as simplivariate models [165] that try to find groups of similarly behaving metabolites. Another route to tackle the problem is to use *a priori* biological information, such as the knowledge of pathways or modules.

Cellular processes arise as the result of many reactions between metabolic intermediates [82]. These reactions are functionally organized in pathways, which together form a large network. Most studies focused on relating changes in pathways to different conditions by using RNA microarray data [31, 59, 102, 204]. Here we describe the extension of a statistical tool, previously developed for analysis of RNA microarray data, to the analysis of metabolomics data.

Studying statistics for a whole group of genes or metabolites avoids the often time-consuming task of multiple testing for each gene or metabolite separately [48]. For metabolomics, predefined groups of pathways [82, 83, 84] or functional modules can be used in this approach. For example, in lipidomics, the test can be performed per lipid class instead of per lipid. Another advantage of group testing is that it can detect differences between conditions that are caused by subtle changes in several metabolites, which are difficult to discover by single metabolite testing.
Nam and Kim [137] distinguished three types of methods for testing pathways, depending on the hypothesis that is tested. The first kind of algorithms test if under particular conditions, a group of genes belonging to a certain pathway is differentially expressed compared with the rest of the genes in the data set (= H1 hypothesis), e.g. T-profiler [14] and PAGE (Parametric Analysis of Gene Set Enrichment) [89]. The second type of methods examines if a selected group of genes from the same pathway has a different behavior under a first condition, compared to a second condition (= H2 hypothesis), e.g. Goeman’s global test [59] and SAM-GS (Significance Analysis of Microarray for Gene Sets) [44]. The third kind of methods, known as Gene Set Enrichment Analysis (GSEA), test the hypothesis that none of the predefined groups of genes in the data set is different between two conditions (= H3 hypothesis). Two types of GSEA are developed: simple GSEA [133, 194] and GSEA using linear models [167, 76]. The tested groups of genes can be predefined groups from e.g. Gene Ontology or KEGG (Kyoto Encyclopedia of Genes and Genomes) pathways [82, 83, 84, 133, 179, 194, 76] or can be defined based on chromosome location [167, 194]. Extensions of GSEA for metabolomics data have been implemented in the web-based tools MSEA (Metabolite Set Enrichment Analysis) [232], MPEA (Metabolite Pathway Enrichment Analysis) [85] and MBRole (Metabolite Biological Role) [29]. In Quantitative Enrichment Analysis (QEA), which is part of MSEA, the Q-statistic from Goeman’s global test was used [232], but the method was not described in the literature about MSEA.

In this chapter, we explain the working of Goeman’s global test for metabolomics in full detail. We discuss the usefulness of this test for establishing significant differences between conditions at the pathway (or module) level. We critically evaluate the validity of the method by using two worked out examples and studying the biological relevance of the test results. For the *E. coli* data set, the test is applied to find pathways that are different under glucose growth compared to acetate growth. With the *S. cerevisiae* data set, the behavior of glycolysis and the tricarboxylic acid (TCA) cycle under three sets of conditions is examined: aerobic versus anaerobic; glucose pulse versus short-term glucose...
Global test for metabolic pathway differences between conditions deprivation (feed off); larger versus smaller glucose pulse. The results show that Goeman’s global test can indeed be used in situations where one wants to know if a metabolic pathway is significantly related to a change in conditions.

3.2 Materials and methods

3.2.1 Escherichia coli data set

GC-MS (Gas chromatography - Mass spectrometry) and LC-MS (Liquid chromatography - Mass spectrometry) data [208] of batch cultures on glucose of E. coli were obtained from TNO Quality of Life (Zeist, The Netherlands). During growth on glucose, acetate is produced. After depletion of glucose, there is a diauxic shift to acetate growth [103]. Sampling of two fermentation processes at eleven time points was performed: four time points in the exponential phase during growth on glucose, five in the post-diauxic phase (growth on acetate), and two in the stationary phase (all carbon sources exhausted) (see Figure 3.1(a)). The data set consists of absolute concentrations (in nanomoles per gram dry weight) of metabolites from glycolysis, the tricarboxylic acid (TCA) cycle and biosynthesis of amino acids, nucleotides and nucleosides. The data are not equidistantly sampled: the time between two subsequent samples ranges from 0.5 to 2 hours. The window of observation is from 10.5 to 20.5 hours elapsed fermentation time (see example for pyruvate, Figure 3.1(b)).
Global test for metabolic pathway differences between conditions

Figure 3.1: (a) Diauxic growth curve. The red points indicate in which phases the measurements were taken. (b) An example of a metabolite concentration profile (pyruvate) under diauxic growth. The different growth phases are indicated on the graph. Abbreviations: nmol, nanomoles; gdwt, gram dry weight.

### 3.2.2 *Saccharomyces cerevisiae* data set

LC-MS data [25, 140, 81] of continuous cultures of *S. cerevisiae* were obtained from the Kluyver Centre for Genomics of Industrial Fermentation (Biotechnology Department, TU Delft, The Netherlands). The cells were cultivated to steady-state in glucose-limited chemostats under aerobic (D=0.1/h) or anaerobic (D=0.05/h) conditions. Furthermore, each steady-state was used to perform a short-term perturbation response experiment, by rapid addition of a concentrated pulse solution and withdrawing samples within a short time frame. Eleven aerobic and four anaerobic experiments were performed. Different perturbations were obtained depending on the composition of the glucose pulse solution. An overview is given in Table 3.1.

---

1chemostat cultures, continuous inflow and outflow
Table 3.1: Description of the experiments.

<table>
<thead>
<tr>
<th>experiment</th>
<th>steady-state condition</th>
<th>perturbation</th>
<th>window of observation (s) (start → end)</th>
<th>time points</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>aerobic</td>
<td>10 mM glucose</td>
<td>0 → 900</td>
<td>13</td>
</tr>
<tr>
<td>2</td>
<td>aerobic</td>
<td>10 mM glucose</td>
<td>0 → 340</td>
<td>13</td>
</tr>
<tr>
<td>3</td>
<td>aerobic</td>
<td>10 mM glucose</td>
<td>0 → 130</td>
<td>13</td>
</tr>
<tr>
<td>4</td>
<td>aerobic</td>
<td>10 mM glucose</td>
<td>0 → 395</td>
<td>13</td>
</tr>
<tr>
<td>5</td>
<td>aerobic</td>
<td>2.5 mM glucose</td>
<td>0 → 454</td>
<td>15</td>
</tr>
<tr>
<td>6</td>
<td>aerobic</td>
<td>2.5 mM glucose</td>
<td>0 → 455</td>
<td>14</td>
</tr>
<tr>
<td>7</td>
<td>aerobic</td>
<td>2.5 mM glucose</td>
<td>0 → 395</td>
<td>14</td>
</tr>
<tr>
<td>8</td>
<td>aerobic</td>
<td>2.3 mM glucose</td>
<td>0 → 118</td>
<td>11</td>
</tr>
<tr>
<td>9</td>
<td>aerobic</td>
<td>2.3 mM glucose + 2.3 mM acetaldehyde</td>
<td>0 → 118</td>
<td>11</td>
</tr>
<tr>
<td>10</td>
<td>aerobic</td>
<td>glucose deprivation (feed off)</td>
<td>0 → 455</td>
<td>14</td>
</tr>
<tr>
<td>11</td>
<td>aerobic</td>
<td>glucose deprivation (feed off)</td>
<td>0 → 455</td>
<td>14</td>
</tr>
<tr>
<td>12</td>
<td>anaerobic</td>
<td>glucose deprivation (feed off)</td>
<td>0 → 216</td>
<td>14</td>
</tr>
<tr>
<td>13</td>
<td>anaerobic</td>
<td>1 mM glucose</td>
<td>0 → 175</td>
<td>14</td>
</tr>
<tr>
<td>14</td>
<td>anaerobic</td>
<td>3 mM glucose</td>
<td>0 → 176</td>
<td>14</td>
</tr>
<tr>
<td>15</td>
<td>anaerobic</td>
<td>3 mM glucose + 3 mM acetaldehyde</td>
<td>0 → 217</td>
<td>14</td>
</tr>
</tbody>
</table>

The data set consists of measurements of absolute metabolite concentrations (in micromoles per gram dry weight) from glycolysis and some of its branches and from the tricarboxylic acid cycle (TCA cycle). The data are not equidistantly sampled: in most experiments the sampling frequency is higher immediately after the pulse and decreases throughout the rest of the time series. The window of observation also differs between experiments.
3.2.3 Data pre-treatment

The intracellular concentrations of different metabolites can differ by more than five orders of magnitude [19]. Furthermore, the abundance of a given compound is not necessarily related to its biological importance [207]. Therefore, the data sets were autoscaled, so that all metabolite levels have zero mean and unit variance. In this way, all compounds are put on the same scale [16].

3.2.4 Goeman’s global test

Assume that \( n \) samples of \( p \) metabolites are measured, of which \( m \) metabolites belonging to the same pathway are selected. Our selection of pathway metabolites is based on the Kyoto Encyclopedia of Genes and Genomes (KEGG) [82, 83, 84]. Let \( i \) be the index for the samples \( (i = \{1, \cdots, n\}) \) and \( j \) the index for the selected metabolites \( (j = \{1, \cdots, m\}) \). The two conditions are labeled with a binary outcome vector \( Y = \{Y_i\} \) of 0’s and 1’s defining the two conditions (e.g. aerobic = 0; anaerobic = 1) (see Figure 3.2(a)). The \((n \times m)\) matrix \( X = (x_{ij}) \) contains concentration levels of selected metabolites. The question do these metabolites behave differently for the two conditions can be translated to the question are the metabolite levels predictive for the outcome. Classically, the method that can be used for this goal is logistic regression [64]. The logistic regression model is defined as [130]:

\[
E (Y_i | \beta) = h^{-1} \left( \alpha + \sum_{j=1}^{m} x_{ij} \beta_j \right)
\]  

(3.1)

where \( \alpha \) is the intercept, \( \beta_j \) the regression coefficient for metabolite \( j \) and \( h \) the logit function [130]:

\[
h (\mu_i) = \ln \left\{ \frac{\mu_i}{1 - \mu_i} \right\}
\]

where \( \mu_i = E (Y_i | \beta) \ (i = \{1, \cdots, n\}) \). The regression coefficients \( \beta_j \) determine the additive effect on the logits of the outcome for a unit
change of metabolite $j$. Stated otherwise, they indicate whether a certain metabolite affects the difference between the two conditions. The regression coefficients $\beta_j$ are all zero if the group of selected metabolites has no influence on the outcome. That answers the question whether this group of metabolites differs between the conditions.

![Figure 3.2: Overview of the score test.](image)

(a) From the autoscaled data matrix, $m$ metabolites belonging to the same pathway are selected. A binary outcome is defined. (b) A score statistic $Q$ is calculated from the mean centered outcome and the matrix of selected metabolites. (c) The significance of the relation between the group of metabolites (pathway) and the outcome is determined by performing a permutation test.

$$P = \frac{\#\text{perm} \geq Q}{\text{total # permutations}}$$

significance level: 0.05

54
Global test for metabolic pathway differences between conditions

The goal is thus to test the null hypothesis $H_0: \beta_1 = \beta_2 = \cdots = \beta_m = 0$ against the alternative hypothesis $H_A$: at least one of the $\beta_j$’s is non zero. Unfortunately, the number of $\beta_j$’s is often much larger than the number of samples leaving no room for classical testing procedures. Goeman [59] states that this problem can be overcome by postulating that all $\beta_j$’s are drawn from some common distribution. This distribution is characterized by $E(\beta) = 0$ and $\text{cov}(\beta) = E(\beta \beta') = \tau^2 I_m$ where $\tau$ is a constant number and $I_m$ the identity matrix of size $m$ [60, 174]. The value $\tau^2$ now regulates the size of the $\beta_j$’s and under the null hypothesis, $E(\beta \beta') = 0$, which means that $\tau^2 = 0$. The null hypothesis and alternative hypothesis are equivalent to testing $H_0 : \tau^2 = 0$ against $H_A : \tau^2 > 0$.

To test the hypothesis, define $r_i = \sum_j x_{ij} \beta_j$ and $r = X \beta$, then $E(r) = 0$ and $\text{cov}(r) = E(rr') = E(X \beta \beta' X') = X E(\beta \beta') X' = X \tau^2 X' = \tau^2 XX'$ [59, 174]. Thus $r$ is a random variable with expectation zero and a covariance containing the parameter $\tau^2$, the same parameter which is relevant in testing for the $\beta_j$’s. By doing this the model in equation 3.1 translates into a random effects model, and the null hypothesis becomes a lack of fit test [108]. This can be tested using Rao’s score test, which has the advantage to be very powerful for detecting small deviations from the null hypothesis [157].

The score statistic is (Figure 3.2(b), see Supplementary Data 1 for a derivation of the statistic) [59]:

$$Q = \frac{(Y - \mu)' R (Y - \mu)}{\mu_2} \quad (3.2)$$

where $\mu = (\mu_1, \cdots, \mu_n)'$ and $\mu_2 = (\mu_{21}, \cdots, \mu_{2n})'$ are the expected value and the variance of $Y$ and $R = \frac{1}{m} XX'$ the configuration matrix of the samples. Under the null hypothesis, $\sum_{j=1}^m x_{ij} \beta_j = 0$, $\mu_i = E(Y_i|\beta) = h^{-1}(\alpha)$ and $\mu_{2i} = \mu_i (1 - \mu_i) (i = 1, \cdots, n)$ for a binary outcome [183]. Because the value of the intercept $\alpha$ is unknown, the exact values of $\mu$ and $\mu_2$ can not be calculated. Therefore, estimates $\hat{\mu}$ and $\hat{\mu}_2$ of the mean and the variance of $Y$ under $H_0$ are used. For a binary outcome $Y$, these estimates are $\hat{\mu}_i = \frac{1}{n} \sum_{i=1}^n Y_i$ and $\hat{\mu}_{2i} = \text{...}$
Global test for metabolic pathway differences between conditions

\[ \hat{\mu}_i (1 - \hat{\mu}_i) (i = 1, \cdots, n) \] [183]. The test statistic becomes:

\[ Q = \frac{(Y - \hat{\mu})' R (Y - \hat{\mu})}{\hat{\mu}^2} \] (3.3)

The statistic \( Q \) is based on differences in (autoscaled) metabolite levels between two conditions (see Figure 3.2(b)). The test statistic is obtained by first multiplying the mean centered outcome with the matrix of selected metabolites (see Figure 3.2(b), step 1). This results in a \((1 \times m)\) vector where each element represents the difference between a sum for the first condition and a sum for the second condition for one of the selected metabolites (Figure 3.2(b), step 1, \( (Y - \mu)' X \) vector). The differences are squared and averaged over the number of selected metabolites \( m \) to obtain a value for the whole group that is not influenced by the number of metabolites (see Figure 3.2(b), step 2). Dividing by the variance of \( Y \) \((= \mu (1 - \mu)\) for a binary outcome) results in a statistic \( Q \) that has a scaled chi square distribution [59] and is therefore statistically more tractable (see Figure 3.2(b), step 3). A p-value for the selected group of metabolites (pathway) is calculated by permuting samples (see Figure 3.2(c)). For small sample sizes, the Q statistic is calculated for all possible permutations. For large sample sizes, the total number of permutations is too large to evaluate them all. Instead, a large number of permutations (for example, 100,000) are used. The p-value is the ratio of the number of times that the Q value of the permuted outcomes is larger or equal than the Q value of the real outcome over the total number of permutations [59]. The p-values were Bonferroni corrected for multiple hypothesis (pathway) testing. A significance level of 0.05 was used for the Bonferroni adjusted p-values.

Goeman’s global test is a method for quickly testing if changes in a pathway are related to different conditions. It detects consistent differences in patterns of metabolite levels between two conditions (see Figure 3.3). It does not test in which direction a pathway is regulated (up or down), nor it determines how many metabolites have changed concentration levels between two conditions.
Figure 3.3: An illustration of the type of differences detected by Goeman’s global test. In this (hypothetical) example, the concentration levels of a group of six metabolites are shown for six samples. Three samples are measured under condition 1 (blue circles) and another three samples under condition 2 (red x-marks). There are consistent differences in the pattern of metabolite levels between the two conditions (metabolites 1, 3, 4 and 5 differ), so the test will give a significant result.

A general problem of testing two conditions, also occurring in biomarker studies [1], is the presence of more than one effect in the data, which increases the variation within one condition. If one effect exceeds the other, this effect has to be removed before the second effect can be tested. This can be done by performing the test on a subset of the data set that is only influenced by the second effect.

3.2.5 Computational tools

Autoscaling was performed by using the PLS Toolbox version 5.2 for use with Matlab [229]. Goeman’s global test was implemented in Matlab [127]. The Matlab code and a short user manual how to use the programs are given in Supplementary Data 5 and 6.
Table 3.2: Classification of the metabolites in the *Escherichia coli* data set.

<table>
<thead>
<tr>
<th>KEGG pathway</th>
<th>number of metabolites in the group</th>
</tr>
</thead>
<tbody>
<tr>
<td>glycolysis + branches</td>
<td>7</td>
</tr>
<tr>
<td>TCA cycle</td>
<td>8</td>
</tr>
<tr>
<td>alanine, aspartate and glutamate metabolism</td>
<td>5</td>
</tr>
<tr>
<td>glycine, serine and threonine metabolism</td>
<td>4</td>
</tr>
<tr>
<td>valine, leucine and isoleucine degradation</td>
<td>4</td>
</tr>
<tr>
<td>valine, leucine and isoleucine biosynthesis</td>
<td>5</td>
</tr>
<tr>
<td>purine metabolism</td>
<td>5</td>
</tr>
</tbody>
</table>

### 3.3 Results and discussion

#### 3.3.1 Results

**3.3.1.1 *Escherichia coli* data set**

The analysis is based on the first nine time points of the two fermentations and compares glucose and acetate growth for different pathways. Because the data are in a quasi steady state during the exponential phase and in another quasi steady state during the post-diauxic exponential phase [227] (see Figure 3.1), they can be treated as non time-resolved data and assigned to two conditions: pre- and post-diauxic shift (glucose versus acetate growth). For each hypothesis, 18 samples are used: 8 of condition 1 (glucose growth) and 10 of condition 2 (acetate growth). In total, there are 26 metabolites used, which can be grouped in seven pathways following the KEGG database [82, 83, 84]. Table 3.2 gives the number of metabolites per hypothesis (tested pathway). The pathways overlap when containing highly connected metabolites (e.g. pyruvate). More information about the overlap can be found in Supplementary Data 2, which gives a list of the metabolites of the *E. coli* dataset, together with their pathway assignments. Table 3.3 reports the Q statistic and the p-value for the different groups. The results show that all studied pathways are different when glucose growth is compared with acetate growth.
Table 3.3: Results of the score test for the *Escherichia coli* data set. The permutation test is based on all permutations. Results are significant when the Bonferroni adjusted p-value is smaller than 0.05. Significant results are indicated in bold.

<table>
<thead>
<tr>
<th>Pathway</th>
<th>Q statistic</th>
<th>not adjusted p-value</th>
<th>Bonferroni adjusted p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>glycolysis + branches</td>
<td>200</td>
<td>$&lt; 10^{-4}$</td>
<td>0.0002</td>
</tr>
<tr>
<td>TCA cycle</td>
<td>135</td>
<td>0.0001</td>
<td>0.0008</td>
</tr>
<tr>
<td>alanine, aspartate and glutamate metabolism</td>
<td>152</td>
<td>$&lt; 10^{-4}$</td>
<td>0.0002</td>
</tr>
<tr>
<td>glycine, serine and threonine metabolism</td>
<td>126</td>
<td>$&lt; 10^{-4}$</td>
<td>0.0002</td>
</tr>
<tr>
<td>valine, leucine and isoleucine degradation</td>
<td>163</td>
<td>$&lt; 10^{-4}$</td>
<td>0.0002</td>
</tr>
<tr>
<td>valine, leucine and isoleucine biosynthesis</td>
<td>163</td>
<td>$&lt; 10^{-4}$</td>
<td>0.0002</td>
</tr>
<tr>
<td>purine metabolism</td>
<td>164</td>
<td>0.0004</td>
<td>0.0027</td>
</tr>
</tbody>
</table>

### 3.3.1.2 *Saccharomyces cerevisiae* data set

For this data set, two groups of metabolites were tested. One was a group of 6 metabolites belonging to glycolysis. The other group exists of 7 metabolites from the TCA cycle. Pyruvate is included in both groups, so in total there are 12 metabolites. A list of metabolites of the *S. cerevisiae* dataset, together with their pathway assignments, is given in the Supplementary Data 3. A test statistic per time point was calculated at four time points: steady state (*t* = 0), 10 seconds, 1 minute and 2 minutes after the pulse. The test was applied to three hypotheses. The first one was if aerobic and anaerobic conditions are different for the selected groups of metabolites. The second hypothesis compared the response to a glucose pulse with the response to short-term glucose deprivation (feed off). The last one looks for differences between larger and smaller glucose pulses (10 mM versus $\leq$ 3 mM). Table 3.4 gives an overview of the number of samples and metabolites used per hypothesis.

The results for the first hypothesis are shown in Table 3.5. At each of the four time points, a significant result (*p* < 0.05) for both groups of metabolites was found when comparing aerobic and anaerobic conditions.

The very large difference between aerobic and anaerobic conditions (large Q statistic, low p-value) makes it difficult to distinguish between
Table 3.4: Overview of the number of samples and metabolites per hypothesis for the *Saccharomyces cerevisiae* data set.

<table>
<thead>
<tr>
<th>outcome pathway number (perturbation 1/ perturbation 2)</th>
<th>pathway</th>
<th>number of metabolites</th>
<th>total number of samples</th>
<th>number of samples of pert.1</th>
<th>number of samples of pert.2</th>
</tr>
</thead>
<tbody>
<tr>
<td>aerobic / anaerobic</td>
<td>glycolysis</td>
<td>6</td>
<td>15</td>
<td>11</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>TCA cycle</td>
<td>7</td>
<td>15</td>
<td>11</td>
<td>4</td>
</tr>
<tr>
<td>glucose pulse (aerobic) / glucose deprivation (aerobic)</td>
<td>glycolysis</td>
<td>6</td>
<td>11</td>
<td>9</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>TCA cycle</td>
<td>7</td>
<td>11</td>
<td>9</td>
<td>2</td>
</tr>
<tr>
<td>glucose pulse of 10 mM (aerobic) / glucose pulse ≤ 3 mM (aerobic)</td>
<td>glycolysis</td>
<td>6</td>
<td>11</td>
<td>4</td>
<td>7</td>
</tr>
<tr>
<td></td>
<td>TCA cycle</td>
<td>7</td>
<td>11</td>
<td>4</td>
<td>7</td>
</tr>
</tbody>
</table>

another set of two conditions. Therefore, the remaining hypotheses were tested only on the aerobic data (see Table 3.4). At steady state (before the perturbation), the experimental conditions are the same for all aerobic experiments (aerobic, glucose-limited, steady state). There are only small fluctuations in metabolite levels due to natural variation and experimental noise. Because the second and the third hypothesis test the effect of different perturbations, they are only tested for the time points at 10 s, 1 min and 2 min (after the perturbation). The results are shown in Table 3.6. The test gave a significant p-value for glycolysis and the TCA cycle for the last two time points when glucose pulse experiments are compared with short-term glucose deprivation (feed off).

For the first time point, a significant p-value for glycolysis was found
Global test for metabolic pathway differences between conditions

Table 3.5: Results of the score test for the *Saccharomyces cerevisiae* data set when comparing aerobic and anaerobic conditions. The permutation test is based on all permutations. Results are significant when the Bonferroni adjusted p-value is smaller than 0.05. Significant results are indicated in bold.

<table>
<thead>
<tr>
<th>pathway</th>
<th>time point</th>
<th>Q statistic</th>
<th>not adjusted p-value</th>
<th>Bonferroni adjusted p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>glycolysis</td>
<td>steady state</td>
<td>114</td>
<td>0.001</td>
<td>0.002</td>
</tr>
<tr>
<td></td>
<td>10 s</td>
<td>81</td>
<td>0.001</td>
<td>0.002</td>
</tr>
<tr>
<td></td>
<td>1 min</td>
<td>79</td>
<td>0.001</td>
<td>0.002</td>
</tr>
<tr>
<td></td>
<td>2 min</td>
<td>83</td>
<td>0.002</td>
<td>0.003</td>
</tr>
<tr>
<td>TCA cycle</td>
<td>steady state</td>
<td>166</td>
<td>0.001</td>
<td>0.002</td>
</tr>
<tr>
<td></td>
<td>10 s</td>
<td>147</td>
<td>0.001</td>
<td>0.002</td>
</tr>
<tr>
<td></td>
<td>1 min</td>
<td>128</td>
<td>0.001</td>
<td>0.002</td>
</tr>
<tr>
<td></td>
<td>2 min</td>
<td>106</td>
<td>0.001</td>
<td>0.002</td>
</tr>
</tbody>
</table>

when comparing large and small glucose pulses.

### 3.3.2 Discussion

Comparing glucose with acetate growth in *E. coli*, our study showed significant differences for all studied pathways. Lowry *et al.*[113] reported that the carbon source (glucose, acetate) affects the intermediates of glycolysis, TCA cycle, purine metabolism and alanine, aspartate and glutamate metabolism. Glycine, serine and threonine metabolism have a precursor in glycolysis (3-phosphoglycerate) (see Supplementary Data 4). Therefore, it is likely that the effect is also observable in these pathways. In the same way the difference between acetate and glucose growth also manifests itself in valine, leucine and isoleucine biosynthesis and degradation (see Supplementary Data 4).

For the *S. cerevisiae* data set, both glycolysis and the TCA cycle are significantly different when aerobic conditions are compared with anaerobic conditions. For glycolysis, an explanation could be that under
Global test for metabolic pathway differences between conditions

Table 3.6: Results of the score test for the 11 aerobic experiments in the *Saccharomyces cerevisiae* data set. The permutation test is based on all permutations. Results are significant when the Bonferroni adjusted p-value is smaller than 0.05. Significant results are indicated in bold.

<table>
<thead>
<tr>
<th>outcome (perturbation 1 / perturbation 2)</th>
<th>pathway</th>
<th>time point</th>
<th>Q statistic</th>
<th>not adjusted p-value</th>
<th>Bonferroni adjusted p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>glucose pulse/glucose deprivation</td>
<td>glycolysis</td>
<td>10 s</td>
<td>18</td>
<td>0.036</td>
<td>0.073</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1 min</td>
<td>53</td>
<td>0.018</td>
<td><strong>0.036</strong></td>
</tr>
<tr>
<td></td>
<td></td>
<td>2 min</td>
<td>51</td>
<td>0.018</td>
<td><strong>0.036</strong></td>
</tr>
<tr>
<td></td>
<td>TCA cycle</td>
<td>10 s</td>
<td>3</td>
<td>0.273</td>
<td>0.545</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1 min</td>
<td>18</td>
<td>0.018</td>
<td><strong>0.036</strong></td>
</tr>
<tr>
<td></td>
<td></td>
<td>2 min</td>
<td>27</td>
<td>0.018</td>
<td><strong>0.036</strong></td>
</tr>
<tr>
<td>glucose pulse of 10 mM/glucose pulse ≤ 3 mM</td>
<td>glycolysis</td>
<td>10 s</td>
<td>31</td>
<td>0.003</td>
<td><strong>0.006</strong></td>
</tr>
<tr>
<td></td>
<td></td>
<td>1 min</td>
<td>11</td>
<td>0.118</td>
<td>0.236</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2 min</td>
<td>10</td>
<td>0.112</td>
<td>0.224</td>
</tr>
<tr>
<td></td>
<td>TCA cycle</td>
<td>10 s</td>
<td>4</td>
<td>0.115</td>
<td>0.230</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1 min</td>
<td>8</td>
<td>0.112</td>
<td>0.224</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2 min</td>
<td>5</td>
<td>0.430</td>
<td>0.861</td>
</tr>
</tbody>
</table>

anaerobic conditions, glycolysis is the principal energy source, and thus of adenosine triphosphate (ATP) [4]. The pyruvate formed by glycolysis is fermented to ethanol [221]. If oxygen is available, pyruvate is further oxidized in the TCA cycle to yield more ATP [4], so that less glucose is needed for supplying the same amount of ATP. Therefore, glucose utilization is more efficient under anaerobic than under aerobic conditions [4], leading to different levels of glycolytic intermediates between aerobic [125] and anaerobic conditions [140].

The TCA cycle and the electron-transport chain have two reactions in common: the succinate dehydrogenase (SDH) reaction and the fumarase (FUM) reaction [28]. Therefore, the electron-transport chain can not function without TCA cycle activity, which makes the TCA cycle an important pathway for ATP production under aerobic conditions [22].
The TCA cycle loses its function in ATP synthesis under anaerobic conditions [117], because the electron-transport chain cannot function without oxygen. TCA cycle activity is therefore much lower in the absence than in the presence of oxygen [28, 22, 117], resulting in different TCA cycle metabolite levels under aerobic compared to anaerobic conditions. For both glycolysis and TCA cycle, a significantly different behavior is observed after a minute when comparing aerobic glucose pulses with glucose deprivation. Glucose removal causes a decrease in fructose-1,6-bisphosphate (FBP), an activator of pyruvate kinase (PYK). Lower FBP levels result in lower PYK activity, which causes phosphoenolpyruvate (PEP) accumulation. PEP inhibits phosphofructokinase (PFK) and represses glycolysis [15]. Glucose deprivation enhances respiration and down regulates fermentation [15, 141, 163], increasing the levels of the TCA cycle intermediates [15, 230].

When larger (10 mM) and smaller (≤ 3 mM) glucose pulses are compared, a significant result for glycolysis is only observed at the beginning of the experiments. This effect is due to the rapid glucose uptake during the first 30 seconds after the pulse [124, 125]. After this period of 30 seconds, the metabolite levels decrease to a steady state [125, 216], which is only slightly higher than the initial steady state [125]. Therefore, the influence of the extent of the pulse is only observable directly (< a minute) after the pulse. The largest part of the glucose influx is redirected to the ethanol branch [231]. As a consequence, a larger pulse has no significant influence on the TCA cycle.

From all the points mentioned above it can be concluded that the results of Goeman’s global test, applied to metabolomics data, correspond with physiology as described in the literature.

When analyzing time points from dynamic experiments, the results of the test changed in time, indicating that Goeman’s global test is able to detect the propagation of perturbations along the network. We believe that it will be useful to extend the test to permit the analysis of time series data, instead of discrete time points. This route will be explored in subsequent work.
Global test for metabolic pathway differences between conditions

To our knowledge, this is the first study where an approach for establishing significant differences between conditions at the pathway (or module) level is applied to metabolomics data. Therefore, implementing other tools from microarray studies that have the same goal for application on metabolomics data and conducting a comparative study of these tools is a direction for future research.

An analysis technique often used in metabolomics is PLS-DA (Partial Least Squares - Discriminant Analysis) [138, 226]. As the name indicates this technique falls into the class of discrimination techniques. It builds a linear regression model to discriminate two given classes from each other, for example healthy versus diseased. It generally makes no use of pathway information, although it can also be applied on a group of selected metabolites of a pathway instead of on the whole data set. Goeman’s global test uses logistic regression and tests if all regression coefficients are zero. Testing at once that not all metabolites in a pathway are identical under two conditions is not a trivial task to perform with PLS-DA. Goeman’s global test provides an easy way to carry out this task with a single test and is therefore a more direct tool to look for changes in pathways.

In this study, Goeman’s global test was applied on microorganisms to relate a pathway with different environmental conditions, but also other applications of this method are possible. In medical biology where one has data of healthy versus diseased people, Goeman’s global test can be used to examine if a certain metabolic pathway is significantly related with having a disease or not. The samples from the healthy people can be regarded as condition 1 and those of the diseased people as condition 2. In the cases that the tested pathway is activated or inhibited by the disease, large differences in metabolite levels between healthy and diseased can be detected. This will result in a large Q statistic and a small p-value. Goeman’s global test can also be performed when some metabolites from the studied pathway are missing due to for example sensitivity of the analytical method, like it was the case for the data sets in this study. However, the results will change, depending on which metabolites are included. If the correlation of the missing metabolite with the outcome is almost equal to the average metabolite-outcome
correlation for the pathway, this has almost no effect on the Q-statistic. If a metabolite that has a much higher or lower correlation to the outcome than average is missing, the value of the Q statistic will decrease or increase respectively.

As mentioned in the methods section, Goeman’s global test is able to detect small deviations from the null hypothesis, which is advantageous when a difference in regulation of a pathway causes only very small differences in concentration levels of only a few metabolites in the studied pathway. In this case Goeman’s global test will also detect the changes in the pathway between two conditions.

As a consequence of the property of detecting small differences between conditions, a significant result can also be obtained when there is a change in the concentration of only one metabolite. This can be illustrated by the following hypothetical experiment. Three data sets of normally distributed random numbers with zero mean and unit variance are generated with 15 rows and 5, 20 and 30 columns respectively. Each of the three data sets is combined with the autoscaled column of steady state concentrations of fructose-1,6-bisphosphate (FBP) from the Saccharomyces cerevisiae data set. The steady state concentrations of FBP differ several orders of magnitude between aerobic and anaerobic conditions. The Q statistic and p-value for testing if there is a difference between aerobic and anaerobic conditions are calculated for the three data sets consisting of the steady state FBP column and the 5, 20 and 30 columns of random data respectively. The results are shown in Table 3.7. Significant results are found when the FBP column is combined with 5 and 20 columns of random numbers. For the data set with the FBP column and 30 columns of random numbers, the result is not significant. The larger the number of random columns added to the data set, the lower the value of the Q statistic and the higher the p-value.
Global test for metabolic pathway differences between conditions

Table 3.7: Results of Goeman’s global test for the data sets consisting of the autoscaled concentrations of FBP in the *Saccharomyces cerevisiae* data set and columns of normally distributed random numbers with zero mean and unit variance. Significant results are indicated in bold.

<table>
<thead>
<tr>
<th>data set</th>
<th>Q statistic</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>FBP + 5 columns of random numbers</td>
<td>41.0395</td>
<td>0.003</td>
</tr>
<tr>
<td>FBP + 20 columns of random numbers</td>
<td>21.3772</td>
<td>0.046</td>
</tr>
<tr>
<td>FBP + 30 columns of random numbers</td>
<td>19.9447</td>
<td>0.081</td>
</tr>
</tbody>
</table>

When a metabolite belongs to more than one pathway, a change in the level of only this metabolite can also mean that another pathway than the one under study is changed between two conditions. In this case, Goeman’s global test can result in a false positive. As an option, one could test the null hypothesis that at least one of the regression parameters is zero ($H_0$: $\beta_1 = 0$ or $\beta_2 = 0$ or $\cdots$ or $\beta_m = 0$) against the alternative hypothesis that all parameters are non-zero ($H_A$: $\beta_1 \neq 0$ and $\beta_2 \neq 0$ and $\cdots$ and $\beta_m \neq 0$). This is the goal of so-called Intersection-Union Tests (IUT) [10, 9, 42], which are not yet generalized for metabolomics. If these tests would be used in metabolomics, they would detect only differences in a group of metabolites (from the same pathway) if the levels of all metabolites in the pathway change. This is often not the case, like it is shown in the following example. The presence of oxygen influences the activity of glycolysis [4]. As can be observed in Figure 3.4, this leads to different levels of only three of the measured metabolites in this pathway (fructose-1,6-bisphosphate, phosphoenolpyruvate and pyruvate). When a null hypothesis like that of IUT would be tested, this would lead to a non-significant result. So these types of tests would be too strict in detecting differences in pathways between conditions, which can result in false negatives.
Global test for metabolic pathway differences between conditions

Figure 3.4: Steady state concentration levels of metabolites of glycolysis in the *S. cerevisiae* data set under aerobic (blue circles) and anaerobic (red x-marks) glucose-limited conditions. Abbreviations: G6P, glucose-6-phosphate; F6P, fructose-6-phosphate; FBP, fructose-1,6-bisphosphate; 3PG, 3-phosphoglycerate; PEP, phosphoenolpyruvate; PYR, pyruvate.

In summary, one can conclude that a good pathway statistic would be a test for a null hypothesis that is intermediate between the $H_0$ of Goeman’s global test and the $H_0$ of IUT (for example, $H_0$: Less than two of the regression parameters are non zero, against the alternative $H_A$: At least two of the regression parameters are non-zero). Developing such tests will be a direction for future research.

Currently, the applicability of tools like Goeman’s global test is restricted by limited network coverage. Data sets contain metabolites from only a limited number of pathways, compared to the whole metabolic network of an organism. This means that only a relatively small number of pathways can be tested. This is expected to improve in the near-future thanks to continuous improvements in analytical technologies for
Global test for metabolic pathway differences between conditions

3.4 Conclusion

Calculating a single statistic for a group of metabolites avoids multiple testing for each metabolite separately. Predefined groups of pathways or functional modules can be used in this approach. The feasibility of using Goeman’s global test, originally designed for microarray data, in metabolomics was studied. To apply Goeman’s global test in metabolomics, the data have to be scaled, because the abundance of a given metabolite is not necessarily related to its biological importance. The results of Goeman’s global test correspond with the physiology of studied organisms, which shows that the test is applicable in metabolomics. In this study the predefined groups were pathways, but the approach can also be extended to functional groups (e.g. lipid classes).

Acknowledgments

This project was financed by the Netherlands Metabolomics Centre (NMC) which is part of the Netherlands Genomics Initiative /Netherlands Organization for Scientific Research. We thank Mariet van der Werf and Peter Punt (TNO, Zeist, The Netherlands) for providing us with the E. coli dataset. We thank Daniël J. Vis (University Medical Centre Utrecht and University of Amsterdam, The Netherlands) for his comments on the manuscript. Gooitzen Zwanenburg (University of Amsterdam, The Netherlands) is gratefully acknowledged for checking the calculations in Supplementary Data 1. We thank Iven Van Mechelen (KU Leuven, Belgium) for giving suggestions for improving the manuscript.

Supplementary Data

The supplementary data is too extensive to be integrally included in this thesis. It can be accessed online at http://dx.doi.org/10.1016/j.aca.2011.12.051
Chapter 4

Inferring differences in the distribution of reaction rates across conditions

Elucidating changes in the distribution of reaction rates in metabolic pathways under different conditions is a central challenge in systems biology. Here we present a method for inferring regulation mechanisms responsible for changes in the distribution of reaction rates across conditions from correlations in time-resolved data. A reversal of correlations between conditions reveals information about regulation mechanisms. With the use of a small *in silico* hypothetical network, based on only the topology and directionality of a known pathway, several regulation scenarios can be formulated. Confronting these scenarios with experi-

---

3This chapter is based on Diana M. Hendrickx, Huub C.J. Hoefsloot, Margriet M.W.B. Hendriks, Daniël J. Vis, André B. Canelas, Bas Teusink and Age K. Smilde. Inferring differences in the distribution of reaction rates across conditions. Molecular Biosystems, Volume 8:9 (2012), pages 2415-2423.
Inferring differences in the distribution of reaction rates across conditions

mental data results in a short list of possible pathway regulation mechanisms associated with the reversal of correlations between conditions. This procedure allows for the formulation of regulation scenarios without detailed prior knowledge of kinetics and for the inference of reaction rate changes without rate information. The method was applied to experimental time-resolved metabolomics data from multiple short-term perturbation-response experiments in S. cerevisiae across aerobic and anaerobic conditions. The method’s output was validated against a detailed kinetic model of glycolysis in S. cerevisiae, which showed that the method can indeed infer the correct regulation scenario.

4.1 Introduction

In living organisms growth, reproduction, maintenance of homeostasis, and adaptation to the environment are enabled by chemical means [185]. The intermediates of metabolic reactions, metabolites, are organized in a large network, consisting of pathways. Elucidating how these pathways function is an important topic in systems biology, because it can contribute to different disciplines [71, 181], such as microbiology [50], plant biology [134, 205] and biomedical sciences [43].

One of the approaches to infer information about the functioning of cellular systems is correlation analysis of metabolomics data [21, 189, 192]. It has been shown that the interpretation of correlations in metabolomics is not straightforward [21, 189], mostly because there is no direct relation between the correlation coefficients and the underlying pathway of reactions [66, 192]. However, interpreting correlations can give more insight in the regulation of biochemical processes [205]. Previous research [21, 189, 192] focused on the origin of correlations in steady state (static) metabolomics data (biological replicates). Table 4.1 gives an overview of how to interpret correlations between metabolites in static data.

Besides correlations, several other approaches have been developed for the reverse engineering of metabolic networks [5, 37]. In a previous study [66], we examined the feasibility of inferring a de novo metabolic network from metabolomics data. The conclusion was that the mathe-
Inferring differences in the distribution of reaction rates across conditions

Table 4.1: Interpretation of correlations in steady state metabolomics data.

<table>
<thead>
<tr>
<th>observation</th>
<th>interpretation</th>
<th>reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>high positive correlation between two metabolites in one condition</td>
<td>(1) dominance of one parameter (e.g. an enzyme concentration)</td>
<td>[21, 189]</td>
</tr>
<tr>
<td></td>
<td>(2) equilibrium</td>
<td>[21, 189]</td>
</tr>
<tr>
<td>negative correlation between two metabolites in one condition</td>
<td>(1) metabolites are not in equilibrium</td>
<td>[21]</td>
</tr>
<tr>
<td></td>
<td>(2) metabolites belong to a moiety conservation</td>
<td>[21]</td>
</tr>
<tr>
<td>sustained correlation across multiple conditions</td>
<td>(1) rapid equilibrium</td>
<td>[189]</td>
</tr>
<tr>
<td></td>
<td>(2) mass conservation</td>
<td>[189]</td>
</tr>
<tr>
<td>reversed correlation between two conditions</td>
<td>(1) change in regulation</td>
<td>[189]</td>
</tr>
<tr>
<td></td>
<td>(2) existence of multiple steady states</td>
<td>[189]</td>
</tr>
</tbody>
</table>

Mathematical methods require sampling frequencies and noise levels that are not in accordance with contemporary laboratory experiments.

This chapter presents a study on the inference of changes in reaction rate distribution from correlations in time-resolved data measured under different conditions, applied to the central carbon metabolism of *Saccharomyces cerevisiae*. The results of the correlation analysis are combined with *a priori* information about the reaction scheme to infer possible regulation scenarios for the pathway. This is facilitated by using an *in silico* hypothetical network model.

Changing conditions are often associated with differences in the distribution of reaction rates [56, 62, 142, 154]. This research shows that correlations in time-resolved metabolomics data, combined with prior knowledge about the pathway, can help to obtain information about changes in the distribution of reaction rates between conditions. This means that correlations in metabolomics time series provide a new computational approach for deriving information about reaction rates, without performing reaction rate analysis. Reaction rates in a metabolic pathway are not straightforward to determine experimentally [39], so
Inferring differences in the distribution of reaction rates across conditions

Table 4.2: Description of the experiments.

<table>
<thead>
<tr>
<th>steady-state condition</th>
<th>perturbation</th>
<th>number of replicates</th>
</tr>
</thead>
<tbody>
<tr>
<td>aerobic</td>
<td>10 mM glucose</td>
<td>4</td>
</tr>
<tr>
<td>aerobic</td>
<td>2.5 mM glucose</td>
<td>3</td>
</tr>
<tr>
<td>aerobic</td>
<td>2.3 mM glucose</td>
<td>1</td>
</tr>
<tr>
<td>anaerobic</td>
<td>1 mM glucose</td>
<td>1</td>
</tr>
<tr>
<td>anaerobic</td>
<td>3 mM glucose</td>
<td>1</td>
</tr>
</tbody>
</table>

the development of computational methods to infer information about reaction rates is very important. Very commonly used methods for the determination of reaction rates are based on measured external fluxes (e.g. metabolic flux analysis [237]). The approach in this chapter makes a link between intracellular metabolite concentrations and intracellular reaction rates by using correlations. This study examines which prior knowledge is needed and how to combine this knowledge with correlations in time-resolved data to detect differences in the distribution of reaction rates across conditions.

4.2 Materials and methods

4.2.1 Data set

The proposed approach was tested on time-resolved metabolomics data from *S. cerevisiae* cultivations. The cells were grown in aerobic \((D = 0.1 \, h^{-1})\) or anaerobic \((D = 0.05 \, h^{-1})\) glucose-limited continuous cultures, as described elsewhere \([25, 67, 81, 140]\) and short-term perturbation-response experiments were performed on the steady state, by introducing a sudden increase in the extracellular glucose concentration (called a glucose pulse). An overview is given in Table 4.2. Quantitative data on absolute levels of intracellular metabolites were obtained by LC-MS/MS, as described elsewhere \([25, 81, 140]\). The metabolites measured are from glycolysis (glucose-6-phosphate, fructose-6-phosphate, fructose-1,6-bisphosphate, 3-phosphoglycerate, phosphoenolpyruvate, pyruvate) and
Inferring differences in the distribution of reaction rates across conditions

the tricarboxylic acid cycle (pyruvate, citrate, oxoglutarate, succinate, fumarate, malate). Each time series consists of 11 time points. The data are plotted in Supplementary Figures 8-10.

4.2.2 Procedure to infer regulation scenarios

The procedure to infer regulation scenarios is depicted in Figure 4.1. Correlation analysis is performed on time-resolved data measured under different conditions (see section Correlation analysis). A priori information about the studied pathway is used to construct an in silico hypothetical network model (see section In silico hypothetical network model). Regulation scenarios for the pathway are inferred by combining the correlations with the hypothetical model.

![Diagram](image)

Figure 4.1: Procedure to infer regulation scenarios from time-resolved data and a priori information about the pathway.

4.2.2.1 Correlation analysis.

For each experiment, the Pearson product-moment correlation [151] between time profiles of metabolite concentrations, consisting of 11 time points is calculated. In the case of replicates, the correlation coefficients are averaged to obtain one statistic for each condition. Attention must be given to the skewed distribution of Pearson’s correlation coefficient, which makes the mean of the correlations not a good measure of central tendency. To calculate averages, the correlations were converted to scores that have a normal distribution, by applying the Fisher z-transform [54]. The mean z-score over replicates over a condition was
Inferring differences in the distribution of reaction rates across conditions

calculated and converted back to a Pearson correlation coefficient by applying the inverse of the Fisher z-transform.

4.2.2.2 *In silico* hypothetical network model.

Based on only the topology and the directionality of the pathway, we can construct a small hypothetical network as follows. The reaction scheme of the studied pathway is retrieved from MetaCyc (YeastCyc) [26]. No prior knowledge about the kinetic parameters is needed. If reactions are described as rapid equilibrium in the literature, substrate and product can be lumped together. Cofactors and branches of the pathway not occurring in the data set are not included in the hypothetical network. Reaction kinetics are formulated using first order mass-action kinetics with all parameters initially set to 1 for simplicity. Regulation scenarios are inferred with the following forward selection procedure:

- **Step 1**: Calculate the steady state before the pulse.

- **Step 2**: Perturb the system from steady state by increasing the inflow. If the qualitative behavior (increase or decrease of metabolite concentrations) after the pulse is in accordance with the experimental data, then it can be concluded that the distribution of the reaction rates does not change after the pulse. If not, then check in the literature if the pathway can have different behavior (e.g. cycle vs. two branches). If different behavior of the pathway is possible, go to Step 3a. If not, go to Step 3b.

- **Step 3a**: Simulate the following scenarios: 1) change of one rate constant (increase or decrease, for a network with \( p \) reactions, there are \( 2p \) scenarios); 2) change of the behavior of the pathway according to literature. When one or more scenarios are in accordance with the observed correlation in the experimental data, then it can be concluded that these are the possible regulation scenarios after the pulse. If not, then select the scenarios most similar to the data (in qualitative behavior). Go to Step 4.

- **Step 3b**: Simulate the following scenario: change of one rate constant (increase or decrease, for a network with \( p \) reactions, there
Inferring differences in the distribution of reaction rates across conditions are $2p$ scenarios. When one or more scenarios are in accordance with the observed correlation in the experimental data, then it can be concluded that these are the possible regulation scenarios after the pulse. If not, then select the scenarios most similar to the data (in qualitative behavior). Go to Step 4.

- Step 4: Change one more parameter in the selected scenarios (per selected scenario, the total number of possibilities is twice the number of remaining parameters). When one or more scenarios are in accordance with the observed correlation in the experimental data, then it can be concluded that these are the possible regulation scenarios after the pulse. If not, then select the scenarios most similar to the data (in qualitative behavior). Repeat Step 4 until scenarios are derived that are in accordance with the experimental data.

The method described above infers the simplest possible regulation scenarios that are in accordance with the changes in correlation observed in the experimental data. The inferred regulation scenarios are based on mass balances and therefore independent of the order of the mass-action kinetics used in the hypothetical network model (see paragraph 3 of the Supplementary Material for an example and a general mathematical proof).

Matlab’s ordinary differential equation solver ode15s [127] was used for calculating concentration values over time and determining steady state concentrations.

4.3 Results

4.3.1 Data set

Figures 4.2, 4.3, 4.4 and 4.5 summarize the results of the correlation analysis for glycolysis and the TCA cycle. The emphasis is put on a qualitative study of metabolite associations, not on the exact values of the correlation coefficient $r$. Therefore the correlations are divided into categories, based on literature [21, 123]: strong ($|r| \geq 0.8$), moderate ($0.6 \leq |r| < 0.8$), weak ($0.3 \leq |r| < 0.6$) and zero ($|r| < 0.3$). The sign
Inferring differences in the distribution of reaction rates across conditions

Figure 4.2: Correlations between measured glycolytic metabolites under aerobic conditions. Abbreviations: GLUC, glucose; ACALD, acetaldehyde; G6P, glucose-6-phosphate; F6P, fructose-6-phosphate; FBP, fructose-1,6-bisphosphate; 3PG, 3-phosphoglycerate; PEP, phosphoenolpyruvate; PYR, pyruvate.

of the correlations is also reported.

The results show metabolite pairs that are positively correlated under all studied experimental conditions: G6P-F6P, 3PG-PEP and FUM-MAL. There were two pairs where no or only weak correlation was observed for both aerobic and anaerobic conditions: 3PG-PYR and PEP-PYR. For the metabolite pairs FBP-3PG, FBP-PEP, CIT-OGL, CIT-FUM and CIT-MAL, the response to a glucose pulse shows negative correlation under aerobic conditions and positive correlation under anaerobic conditions. Under aerobic conditions, positive correlation for the 10 mM glucose pulses and zero correlation for the 2.3-2.5 mM glucose
Inferring differences in the distribution of reaction rates across conditions

1 - 3 mM glucose anaerobic

![Diagram of glycolytic metabolites](image)

- **strong positive** $(r \geq 0.8)$
- **moderate positive** $(0.6 \leq r < 0.8)$
- **low positive** $(0.3 \leq r < 0.6)$
- **very low** (zero) $(-0.3 < r < 0.3)$
- **low negative** $(-0.6 < r \leq -0.3)$
- **moderate negative** $(-0.8 < r \leq -0.6)$
- **strong negative** $(r \leq -0.8)$

Figure 4.3: Correlations between measured glycolytic metabolites under anaerobic conditions. Abbreviations: see Figure 4.2.

Pulses is observed for SUC-MAL and SUC-OGL. Under anaerobic conditions, responses to glucose pulses are negatively correlated for these two metabolite pairs. The sustained correlations between the metabolite pairs G6P-F6P and MAL-FUM in the time-resolved data correspond to results from previous studies of correlations in steady-state data and point to rapid equilibrium reactions [62, 189]. From the preserved correlation of 3PG-PEP we can infer that both reactions from 3PG to 2PG and from 2PG to PEP, catalyzed by phosphoglyceromutase and enolase respectively, have to be rapid equilibrium reactions, which is in accordance with the literature [35]. Furthermore, in both steady state and time-resolved data, there is a relation between the displacement from equilibrium of a reaction and the strength of the correlation between substrate and product. The closer the reaction is to its equilibrium, the stronger the correlation between substrate and product [21].

The reversed correlation between aerobic and anaerobic conditions for some of the metabolite pairs in glycolysis and the TCA cycle were unexpected. A small hypothetical network model was used to determine the regulation mechanisms that could explain this behavior.
Inferring differences in the distribution of reaction rates across conditions

Figure 4.4: Correlations between measured metabolites of the TCA cycle under aerobic conditions. Abbreviations: PYR, pyruvate; CIT, citrate; OGL, oxoglutarate; SUC, succinate; FUM, fumarate; MAL, malate.
4.3.2 Hypothetical network model

4.3.2.1 Glycolysis.

To examine which information about the network can be inferred from the reversed correlation between FBP and 3PG, the hypothetical network model of Figure 4.6 can be used. This hypothetical network is obtained as follows. G6P and F6P are in rapid equilibrium (see results correlation analysis) and are lumped together. The equilibrium pool of GAP and DHAP [201] is presented by one node. The pair of reactions $\text{GAP} \rightarrow \text{BPG}$ and $\text{BPG} \rightarrow \text{3PG}$ is lumped together because BPG is not measured. The part of glycolysis after 3PG is left out because this part of glycolysis is not relevant to derive regulation scenarios from the negative correlation between FBP and 3PG. FBP and 3PG correspond to $B$ and $D$ in the hypothetical network model. Under aerobic conditions, a negative correlation between FBP and 3PG in the first few minutes after the perturbation is observed in the data because FBP increases and 3PG decreases (See Supplementary Figures 9(a-d) and 10(a-d)). Based on mass balances, an increase in $B$ and a decrease in $D$ as an initial response to the pulse is only possible in the following situations:

- a smaller rate of the reaction converting $B$ to $C$ (a lower value for $k_2$);
Inferring differences in the distribution of reaction rates across conditions

- a larger rate of the reaction converting $C$ to $E$ in the branch (a higher value for $k_3$);
- a smaller rate of the reaction converting $C$ to $D$ (a lower value for $k_4$);
- a larger rate of the reaction that further metabolizes $D$ (a higher value for $k_6$);

Thus by reasoning, regulation scenarios corresponding to the reversal of correlations between conditions can be inferred. In the case of a more complex network, there is a need for a more automated way to infer regulation scenarios. This can be done by using first order mass-action kinetics for the hypothetical model and performing simulations. The mass balances for the metabolites are

\[
\frac{d[A]}{dt} = in - k_1 * [A] \\
\frac{d[B]}{dt} = k_1 * [A] - k_2 * [B] \\
\frac{d[C]}{dt} = k_2 * [B] - (k_3 + k_4) * [C] \\
\frac{d[D]}{dt} = k_4 * [C] - k_6 * [D] \\
\frac{d[E]}{dt} = k_3 * [C] - k_5 * [E]
\]

with parameter values $in = 1$ and $k_1 = k_2 = k_3 = k_4 = k_5 = k_6 = 1$. The steady state of this system is $[A] = [B] = 1$ and $[C] = [D] = [E] = 0.5$. A pulse, a sudden increase in the inflow, is simulated by changing the inflow to $in = 2$. When the pulse does not cause relative changes in reaction rates, all metabolite concentrations increase (see Figure 4.7). This is not in accordance with the behavior of the experimental data under aerobic conditions. Supplementary Table 1 shows the results of the simulations when one parameter is changed. There are four regulation scenarios leading to an increase in $B$ and a decrease in $D$ as an initial
Inferring differences in the distribution of reaction rates across conditions

response to the pulse (see Figure 4.8), which are the same as those found by reasoning.
Correlation analysis of time-resolved data under different conditions is a valuable method to infer regulation scenarios. The simplest possible scenarios leading to the increase or decrease of metabolite concentrations are considered. The ones that fit the qualitative behavior of the experimental data are scenarios that could be true regulation. This means that a simple first order network model can be used if the pathway is too complex to derive regulation by reasoning. This is also shown for the TCA cycle in paragraph 4.3.2.2.

This example illustrated how the possible regulation scenarios can be inferred from the reversal of correlations under aerobic conditions. These regulation scenarios do not depend on the kinetics (e.g. first or second order kinetics) used in the hypothetical network model. Paragraph 3 of the Supplementary Material shows that the initial qualitative behavior of the metabolite concentration profiles after the pulse is independent of the order of the mass-action kinetics.

If the pulse does not cause relative changes in reaction rates, the response shows positive correlation between B and D (see Figure 4.7). This is what happens in glycolysis after a glucose pulse under anaerobic conditions.

4.3.2.2 TCA cycle.

The hypothetical network model in Figure 4.9 was used to infer possible regulation scenarios from the reversed correlations between metabolites in the TCA cycle. To obtain the hypothetical network, the following pairs of reactions were lumped together because their common intermediate is not measured: PYR → ACoA and ACoA → CIT (k₁); CIT → ISOCIT and ISOCIT → OGL (k₂); OGL → SUCCoA and SUCCoA → SUC (k₃); PYR → OAA and OAA → MAL (k₇). The TCA cycle can show two types of behavior. It plays a crucial role in ATP production under respiratory conditions (for example glucose-limited aerobic conditions) and behaves as a cycle. Under fermentative (for example anaerobic) conditions, TCA cycle activity has only a role in the formation of precursors of amino acids and does not act as a cycle, but as two
Inferring differences in the distribution of reaction rates across conditions

Figure 4.6: *In silico* hypothetical network model used to study reversal of correlations in glycolysis. Abbreviations: G6P, glucose-6-phosphate; F6P, fructose-6-phosphate; FBP, fructose-1,6-bisphosphate; GAP, glyceraldehyde-3-phosphate; DHAP, dihydroxyacetone phosphate; BPG, 1,3-bisphosphoglycerate; 3PG, 3-phosphoglycerate; GLYC, glycerol.

Figure 4.7: Response of the metabolites in the *in silico* hypothetical network model of Figure 4.6 to an increase in the inflow. The numbers on the axes are in arbitrary units.
Inferring differences in the distribution of reaction rates across conditions

Figure 4.8: Changes in the parameters that lead to the emergence of a negative correlation between B and D: a) a lower value for $k_2$; b) a higher value for $k_3$; c) a lower value for $k_4$; d) a higher value for $k_6$. The numbers on the axes are in arbitrary units.
Inferring differences in the distribution of reaction rates across conditions

branches, leading to oxoglutarate and succinate [22, 161]. Both types of TCA cycle behavior have to be present in the model and prior knowledge about the way the pathway acts before the perturbation is necessary to apply the procedure depicted in Figure 4.1. The mass balances for this system are (see Figure 4.9 for an explanation of the symbols):

\[
\begin{align*}
\frac{d[A]}{dt} &= \text{in} - (k_1 + k_7 + k_8) \times [A] \\
\frac{d[B]}{dt} &= k_8 \times [A] - k_9 \times [B] \\
\frac{d[C]}{dt} &= k_7 \times [A] - (k_{5b} + k_6) \times [C] + k_{5a} \times [E] \\
\frac{d[D]}{dt} &= k_1 \times [A] + k_6 \times [C] + k_2 \times [D] \\
\frac{d[E]}{dt} &= k_{5b} \times [C] - (k_{4b} + k_{5a}) \times [E] + k_{4a} \times [G] \\
\frac{d[F]}{dt} &= k_2 \times [D] - (k_3 + k_{11}) \times [F] \\
\frac{d[G]}{dt} &= k_{4b} \times [E] - (k_{4a} + k_{10}) \times [G] + k_3 \times [F]
\end{align*}
\]

Parameter values are: \( k_{4b} = k_{5b} = k_7 = 0 \) and all other parameters are 1 when the pathway acts as a cycle (for example aerobic glucose-limited conditions); \( k_3 = k_{4a} = k_{5a} = k_6 = 0 \) and all other parameters are 1 when the pathway acts as two branches (for example anaerobic conditions).

When the pathway acts as a cycle, the steady state of this system is \([A] = [B] = 1/2; [C] = [E] = [G] = 1/6; [D] = 2/3; [F] = 1/3\). A pulse is simulated by changing the inflow to \( \text{in} = 2 \). If the pulse did not cause relative changes in reaction rates, all of the metabolite concentrations would increase (see Supplementary Figure 4). In the experimental data, citrate decreases and the other metabolites increase (see Supplementary Figures 9(e-h) and 10(e-h)), which results in the emergence of negative correlations between metabolites in the TCA cycle (CIT-OGL, CIT-MAL and CIT-FUM). A decrease in D and an increase of A, C, E, F and
Inferring differences in the distribution of reaction rates across conditions

G are observed for the model in the following cases (see Supplementary Table 2):

- After the pulse, the pathway acts as two branches (switch from respiration to fermentation) (see Supplementary Figure 5a).

- After the pulse, there is a larger rate of the reaction converting citrate($D$) to oxoglutarate($F$) (higher value for $k_2$, see Supplementary Figure 5b).

The first scenario, the TCA cycle has different behavior before and after the pulse, is in accordance with results of experiments from Kresnowati et al. [104].

When the pathway acts as two branches (for example in the anaerobic case), the following steady state is found: $[A] = [B] = [C] = [D] = [E] = [F] = [G] = 1/3$. A pulse is simulated by changing the inflow to $in = 2$. All concentration profiles would increase if the pulse did not cause relative changes in reaction rates (see Supplementary Figure 6). In the experimental data, pyruvate increases and citrate, malate, fumarate and oxoglutarate decrease (see Supplementary Figure 8(c-d)). Succinate increases and is negatively correlated with citrate, malate, fumarate and oxoglutarate. Changing only one reaction rate did not result in a scenario that is in accordance with the data (see Supplementary Table 3). Decreasing $k_1$ agreed best with the experimental data. Therefore all scenarios where $k_1$ is decreased and a second parameter changed (increase or decrease) were examined. Supplementary Table 4 shows that none of these scenarios is fully consistent with the experimental data. The two scenarios that show the most similarities with the experimental data are: 1) a decrease in $k_1$ and an increase in $k_{4b}$; 2) a decrease in $k_1$ and an increase in $k_{5b}$. All scenarios that are combinations of the former two and a change of a third parameter were simulated. The results are shown in Supplementary Table 5. An increase of A and G and a decrease of C, D, E and F is observed in the following situations:

- a smaller rate from pyruvate($A$) to citrate($D$), a larger rate from fumarate($E$) to succinate($G$), a larger rate from malate($C$) to fumarate($E$) (smaller value for $k_1$, higher values for $k_{4b}$ and $k_{5b}$, see Supplementary Figure 7a);
Inferring differences in the distribution of reaction rates across conditions

Figure 4.9: *In silico* hypothetical network model used to study reversal of correlations in the TCA cycle. Blue arrows: reaction included only in the TCA cycle under respiration (cycle); red arrows: reaction included only in the TCA cycle under fermentation (two branches). Reaction scheme based on Camarasa et al. [22]. Abbreviations: PYR, pyruvate; ETH, ethanol; OAA, oxaloacetate; MAL, malate; CIT, citrate; ISOCIT, isocitrate; FUM, fumarate; OGL, oxoglutarate; SUC, succinate.

- a smaller rate from pyruvate(A) to citrate(D), a larger rate from fumarate(E) to succinate(G), a smaller rate from pyruvate(A) to malate(C) (smaller values for $k_1$ and $k_7$, higher value for $k_{4b}$, see Supplementary Figure 7b);

### 4.4 Validation

The proposed method was validated for glycolysis by examining whether one of the regulation scenarios inferred from the correlation analysis is in accordance with the way glycolysis is regulated after a glucose pulse under aerobic conditions. For this purpose we used the latest detailed kinetic model for aerobic glycolysis in *S. cerevisiae*, developed by van Eu-
Inferring differences in the distribution of reaction rates across conditions

nen [209](see Supplementary Material for details on the model). Figure 4.10 shows the rate of the reactions from FBP to GAP/DHAP ($v_{ALD}$), from DHAP to glycerol ($v_{GLY}$), from BPG to 3PG ($v_{PGK}$) and from 3PG to 2PG ($v_{PGM}$) under aerobic conditions, simulated with the model. The rates of the reactions from FBP to GAP/DHAP and from BPG to 3PG are increasing (see Figure 4.10a and c), and rate of reactions from DHAP to glycerol is constant (due to the model’s assumption of constant rates to the branches, see Supplementary Material and Figure 4.10b). This disproves the first three scenarios proposed in paragraph 4.3.2.1. Figure 4.10d confirms that the fourth glycolysis-case regulation scenario, a larger rate of the reaction from 3PG to 2PG, is the correct one. Larger reaction rates in lower glycolysis after a glucose pulse under aerobic conditions can be explained by the feed-forward activation of pyruvate kinase by FBP, which increases rapidly after the pulse [79] (see Figure 4.11). Because of the feed-forward activation, the reactions in the lower part of glycolysis become faster. As a consequence, the concentration of 3PG decreases and the correlation between FBP and 3PG is negative. More details are given in paragraph 2 of the Supplementary Material.

It can thus be concluded that when no comprehensive model about the studied pathway is available, the possible regulation scenarios for a given pathway can be inferred from correlations in time-resolved data with the help of a simple hypothetical model.

4.5 Discussion

In this study, a correlation-based method for inferring regulation scenarios from time-resolved metabolomics data is illustrated on the central carbon metabolism of S. cerevisiae. The scenarios are inferred by changing one or more parameters in a hypothetical network that is a simplification of the studied pathway. An important advantage of this method is that information about regulation can be obtained without detailed knowledge of kinetic parameters. Furthermore, the method uses straightforward mathematics and is easy to implement.

A major limitation of the method is that the number of scenarios that need to be tested increases with the number of reactions in the hypothe-
Inferring differences in the distribution of reaction rates across conditions

tical network model. In the worst case, $3^n$ scenarios need to be tested for a network with $n$ reactions (every parameter can increase, decrease or remain unchanged). This can be avoided by using the forward selection procedure described in paragraph 4.2.2.2. For example, for the TCA cycle under anaerobic conditions, the reaction scheme consists of nine reactions, so in principle $3^9 = 19683$ scenarios have to be tested. By using the forward selection procedure, only 62 scenarios had to be checked (see Supplementary Tables 3-5).

A drawback of the forward selection method is that added parameters can not be removed later, which can lead to local solutions [122].

Applicability of methods based on time-resolved data is currently restricted by the measured pathways, which is usually only a small part of the metabolism. Often not all metabolites of a pathway are measured. When the common intermediates of reactions are not measured, the reactions can be lumped together.

Because analytical methods continuously improve, it is expected that future time-resolved metabolomics data will cover larger parts of metabolism.
Inferring differences in the distribution of reaction rates across conditions

Figure 4.10: Reaction rates before and after a 2.3 mM glucose pulse under aerobic conditions, simulated with the model of van Eunen [209]. a) rate of the reaction from FBP to GAP/DHAP ($v_{ALD}$) b) rate of reactions from DHAP to glycerol ($v_{GLY}$) c) rate of the reaction from BPG to 3PG ($v_{PGK}$) d) rate of the reaction from 3PG to 2PG ($v_{PGM}$). Abbreviations: ALD, aldolase; GLY, glycerol; PGK, 3-phosphoglycerate kinase; PGM, phosphoglycerate mutase. From the scenarios described in paragraph 4.3.2.1, only the fourth one is in accordance with these figures. Figures a-c disprove the first three scenarios. We can conclude that an increase in the reaction rate that further metabolizes 3PG (to 2PG) is the correct one.
Inferring differences in the distribution of reaction rates across conditions

Figure 4.11: Feed-forward activation of pyruvate kinase by FBP. Abbreviations: GLUC, glucose; G6P, glucose-6-phosphate; F6P, fructose-6-phosphate; FBP, fructose-1,6-bisphosphate; DHAP, dihydroxyacetone phosphate; GAP, glyceraldehyde-3-phosphate; PEP, phosphoenolpyruvate; PYR, pyruvate; PYK, pyruvate kinase.

4.6 Conclusion

In this study, a priori information about the topology and directionality of a pathway is combined with time-resolved data of metabolites to infer possible regulation mechanisms of the network. By combining correlations in time-resolved metabolomics data with a simple hypothetical network model, possible regulation scenarios could be formulated without a priori knowledge of detailed kinetics. Method validation with a detailed kinetic model as benchmark confirmed that the proposed method was indeed able to infer the correct regulation scenario in aerobic glycolysis of S. cerevisiae after a glucose pulse from the correlations in time-resolved metabolomics data.
Inferring differences in the distribution of reaction rates across conditions

Acknowledgments

This project was financed by the Netherlands Metabolomics Centre (NMC) which is part of the Netherlands Genomics Initiative/Netherlands Organization for Scientific Research. Karen van Eunen and Barbara Bakker (University Medical Center Groningen) are gratefully acknowledged for providing us with the simulation models.

Supplementary Data

The supplementary data is too extensive to be integrally included in this thesis. It can be accessed online at http://www.rsc.org/suppdata/mb/c2/c2mb25015b/c2mb25015b.pdf
Inferring differences in the distribution of reaction rates across conditions
In living organisms, cells have mechanisms that enable them to adapt to the environment, while maintaining certain essential functions for survival. Elucidating the principles of cellular adaptation is one of the major challenges in systems biology. Knowledge of cellular adaptation can contribute to different disciplines, such as microbiology, plant biology and medicine.

For a comprehensive study of the functioning of cellular systems, we need to understand their dynamics. Dynamical properties can be studied with kinetic models. However, the development of kinetic models is hampered by the lack of sufficient information about kinetic properties. Therefore, there is a need for mathematical frameworks that use a mini-

---

4Manuscript in preparation: Integrating time-resolved metabolomics data into dynamic flux balance analysis
Integrating time-resolved metabolomics data into dynamic flux balance analysis

mal amount of kinetic information. One of these frameworks is dynamic flux balance analysis (DFBA), a method based on the assumption that cellular systems have evolved towards optimality against perturbations. DFBA has some limitations. The large amount of variables to estimate and the computational complexity make the method less suitable for larger systems. Furthermore, DFBA assumes quasi steady state for the intracellular environment, which makes DFBA not an appropriate method to study highly dynamic responses to perturbations.

Here, we present a mathematical framework to tackle some of the limitations encountered in DFBA. Time-resolved metabolomics data were integrated in the DFBA to reduce both the number of variables to be estimated and the computational complexity, which makes the DFBA more suitable for larger systems. Furthermore, by integration of experimental data with DFBA, dynamics can also be studied in situations where no quasi steady state can be assumed for the intracellular environment.

The mathematical framework presented in this chapter can be used in two ways. Given prior information about the rates through the pathway(s), hypotheses about cellular functioning can be tested. When the optimality principles underlying adaptation to a certain condition are known, DFBA combined with time-resolved data can be used to estimate reaction rate profiles.

Here, a case study of using the mathematical framework for hypothesis testing is presented. The responses to a glucose pulse under aerobic conditions of 5 external and 28 internal metabolites of the central carbon metabolism of *S. cerevisiae* are integrated in DFBA to test hypotheses about optimal cellular behavior.

5.1 Introduction

Under changing environmental conditions, cells are capable to adapt to these changes while maintaining essential functions for survival [90, 187]. Understanding the principles of cellular adaptation is one of the major challenges in systems biology [7, 61, 105, 152]. When the availability of substrates in the environment changes, adjust-
Integrating time-resolved metabolomics data into dynamic flux balance analysis

ment occurs in the distribution of reaction rates of metabolic pathways [149, 90]. It is assumed that the response of cellular systems has evolved towards optimal behavior with respect to these perturbations [90, 152]. To better understand cellular functioning and cellular evolution it is important to analyze reaction rates and explore hypotheses about optimality [139, 90]. Optimality principles, for example maximizing growth or energetic efficiency, can be mathematically described by objective functions [55]. It has been proposed in the literature that there are conditions for which the optimality principles can not be described by a single objective, but by a combination of multiple objectives [176].

External reaction rates can be measured by enzymatic methods [197]. Intracellular reaction rates can not be measured directly, but have to be estimated computationally from other types of biological data [139, 129]. The two most common frameworks for estimating reaction rates are $^{13}$C metabolic flux analysis ($^{13}$C MFA) and flux balance analysis (FBA).

In $^{13}$C MFA, the intracellular reaction rates are estimated from extracellular reaction rates and the concentrations of $^{13}$C-labeled metabolites [228]. $^{13}$C labeling of the medium substrate(s) leads to propagation of labeled carbon atoms through the network of metabolic reactions [212]. The resulting labeling states ($2^n$ for a metabolite with $n$ carbon atoms), also called isotopomers, are measured with GC-MS or LC-MS [144] and the reaction rates are calculated from the labeling data and the external reaction rate measurements [228].

FBA is based on stoichiometry and additional constraints, and does not require detailed knowledge of kinetic parameters [147]. The reaction stoichiometry is retrieved from a database (e.g. KEGG [82, 83, 84]). The mass balances can then be formulated: $dC/dt = S \cdot v$, where $dC/dt$ represent the time derivatives of the metabolite concentrations, $S$ is the stoichiometric matrix and $v$ the flux vector [74]. At steady state, there are no concentration changes and $S \cdot v = 0$ [55]. Since there are generally more reactions than metabolites [214], the system $S \cdot v = 0$ is underdetermined. Therefore, additional constraints have to be imposed. This can be thermodynamic, maximum flux capacity or regulatory constraints [36]. The constraints are based on prior knowledge extracted from the literature or databases. The biochemical goal of the system
Integrating time-resolved metabolomics data into dynamic flux balance analysis (e.g. maximize growth) is written as an objective function that has to be maximized or minimized [55]. The steady state flux distribution is determined by optimizing the objective function, given the constraints. Dynamic FBA (DFBA) is an extension of FBA, developed by Mahadevan et al. [120], that accounts for dynamic changes in the cellular metabolism. Optimization is performed over the entire time period by defining an objective function and dynamic mass balance constraints \( (dC/dt = S \cdot v) \) [120]. Reaction rate profiles are calculated by solving the optimization problem.

In DFBA, both the time derivatives of the metabolite concentrations and reaction rates are unknown [120], which results in a large number of unknowns for larger systems. Furthermore, the differential equation constraints make the optimization problem mathematically complex. Therefore, previous DFBA studies focused on small systems [116, 120, 238] or on the long-term (hours scale) extracellular dynamics, while assuming a quasi steady state for the intracellular metabolism [68, 78].

Here, we present a mathematical framework to solve some of the issues encountered in DFBA. Time-resolved data of metabolite concentrations are integrated in the DFBA to reduce the complexity of the optimization problem. By calculating the time derivatives of the metabolite concentrations from the experimental data and substituting these quantities for the left hand side of the mass balances, the differential equations are converted to linear equations. Incorporating experimental data also reduces the solution space, because the number of unknown variables is reduced. The reduced complexity and the smaller number of unknowns makes the DFBA framework described in this chapter appropriate for larger systems and situations where quasi steady state cannot be assumed for the internal metabolism (e.g. response to short-term perturbations, which is highly dynamic [25]).

The mathematical framework presented in this study can be applied in two ways (see Figure 5.1): for hypothesis testing and for estimating reaction rates. Hypotheses about optimal behavior under perturbations can be formulated as objective functions and serve as input for the DFBA, together with time-resolved experimental data. For each objective, the output of the DFBA (reaction rate profiles) is compared with external
Integrating time-resolved metabolomics data into dynamic flux balance analysis

information about the rates through the pathway(s) to decide which hypotheses can be rejected. In the second application, the external information is used as input to improve reaction rate estimations. If the biological goal of the system is known under the given condition, the corresponding objective function can be used as input for the DFBA, together with the time-resolved data and external information, to estimate reaction rates over time. The external information can be (among others): reaction rates determined with $^{13}$C metabolic flux analysis, predictions from kinetic models, physiological information.

![Diagram](image)

Figure 5.1: Applications of DFBA combined with time-resolved metabolomics data. (a) testing hypotheses (objectives) (b) estimating reaction rates.

With the aim to test hypotheses on optimal cellular behavior, we present a case study of using DFBA combined with time-resolved experimental data. Metabolite concentration profiles resulting from short-term perturbation-response experiments (glucose pulses under aerobic conditions) for *S. cerevisiae* are integrated in the DFBA to test different plausible objective functions. The outcome is compared with external information from physiology and kinetic models.
5.2 Materials and methods

5.2.1 Mathematical framework for integrating time-resolved metabolomics data into DFBA

The mathematical framework for integrating time-resolved metabolomics data in DFBA is depicted in Figure 5.2.

Steady state fluxes are calculated from a genome scale model. Time-resolved metabolomics data are combined with stoichiometric information from the genome scale model to formulate a dynamic optimization problem. Solving the optimization problem results in optimal dynamic reaction rate profiles, together with their upper and lower bounds. In the case of testing hypotheses (objective functions), the reaction rates
Integrating time-resolved metabolomics data into dynamic flux balance analysis

are plotted and compared with external information from the literature to decide which hypotheses can be rejected. The details of the different steps in the mathematical framework are described below.

5.2.1.1 Flux balance analysis (FBA)

Flux balance analysis [147] is applied on a genome scale model. Fluxes of reactions supposed not to occur under the given experimental conditions (e.g. because a certain substrate is not in the medium), are equated to zero. A realistic convex objective function $Z$ for the given conditions is chosen. The reaction rates are constrained by upper and lower bounds, where the lower bound and the upper bound have the same sign for irreversible reactions and different sign for reversible reactions. An optimization problem of the following form is solved:

\[
\begin{align*}
\text{minimize or maximize } & \quad Z \\
\text{subject to} & \quad S \cdot v = 0 \quad \text{(steady state mass balances)} \\
& \quad \text{for each reaction rate } v_j (j = 1, \ldots, r) : \; v_{j,\text{min}} \leq v_j \leq v_{j,\text{max}}
\end{align*}
\]

where $S$ is the stoichiometric matrix of the genome scale model, $v$ is the flux vector and $r$ is the number of reactions in the model. Solving the optimization problem results in optimal steady state fluxes and an optimal value $Z_{opt}$ for the objective function.

5.2.1.2 Flux variability analysis (FVA)

Because the optimization problem described above is convex, the optimal value $Z_{opt}$ is a global optimum [114], but the optimal fluxes in the FBA solution are not always unique. It is possible that the same optimal value $Z_{opt}$ is achieved by different flux distributions satisfying the constraints. Flux variability of the optimal solutions can be studied with Flux Variability analysis (FVA) [121]. For each flux $v_j (j = 1, \ldots, r; \; r = \text{number of reactions in the model})$, the following optimization problems
Integrating time-resolved metabolomics data into dynamic flux balance analysis

are solved:

\[
\begin{align*}
\text{minimize } & \quad v_j \\
\text{subject to } & \quad S \cdot v = 0 \\
& \quad \text{for each reaction rate } v_j (j = 1, \cdots, r) : v_{j,\text{min}} \leq v_j \leq v_{j,\text{max}} \\
& \quad Z = Z_{\text{opt}} \\
\end{align*}
\]

\[
\begin{align*}
\text{maximize } & \quad v_j \\
\text{subject to } & \quad S \cdot v = 0 \\
& \quad \text{for each reaction rate } v_j (j = 1, \cdots, r) : v_{j,\text{min}} \leq v_j \leq v_{j,\text{max}} \\
& \quad Z = Z_{\text{opt}} \\
\end{align*}
\]

In this way lower and upper bounds are obtained for the optimal solutions.

5.2.1.3 Calculation of derivatives

To reduce noise, the data were smoothed with B-splines (piecewise polynomials) [64] and the time derivatives of the spline functions were calculated. The derivatives of the splines are evaluated at equidistant time points which are within the window of observation of all time series in the data set. When a metabolite is measured multiple times, the average of the derivatives is calculated. A \((m \times n)\)-matrix \(dC/dt\) \((m\ \text{metabolites, } n\ \text{time points})\) of time derivatives of the metabolite concentrations \(C\) is obtained.

5.2.1.4 Lumping of reactions

Reactions in the genome scale model that have one or more of the measured metabolites as substrate or product are selected. If there are two reactions \(A \overset{v_1}{\rightarrow} B \overset{v_2}{\rightarrow} C\), and \(B\) is not measured, the reactions are lumped to \(A \overset{v_3}{\rightarrow} C\) where \(v_3 = v_2 = v_1\) (at steady state). Two reactions \(A \overset{v_3}{\rightarrow} B\)
Integrating time-resolved metabolomics data into dynamic flux balance analysis

and $C \xrightarrow{v_3} B$ with $A$ and $C$ not measured are combined to one arrow, other pathways $v_3 \leftrightarrow B$, where $v_3 = v_1 + v_2$.

### 5.2.1.5 Dynamic FBA

Dynamic FBA is an extension of FBA that accounts for dynamics. The dynamic mass balance constraints are of the form $dC/dt = S \cdot v$. When using both external and internal metabolites, it has to be taken into account that the extracellular volume is 100 to 1000-fold higher than the intracellular volume [24]. Therefore, the left-hand side of the mass balances for the external metabolites is multiplied by a factor $V_{ext}/V_{int}$ [34], where $V_{int}$ and $V_{ext}$ are the intracellular and extracellular volume respectively.

In previous studies [109, 116, 91, 120], the differential equations are rewritten as algebraic equations using a finite collocation method. For larger systems, this takes an unreasonable amount of computational time [120]. When measurements are available, the differential equations can be rewritten as algebraic equations by substituting the derivatives calculated from data (see 5.2.2.3.) in the differential equations. In this way the differential equation constraints are converted into linear constraints.

Put $A = \begin{pmatrix} S_l & 0 & \cdots & 0 & 0 \\ 0 & S_l & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & S_l & 0 \\ 0 & 0 & \cdots & 0 & S_l \end{pmatrix}$ and $b = \text{vec}(dC/dt)$, where $S_l$ is the stoichiometric matrix of the lumped model and $dC/dt$ are the time derivatives calculated from the measured data. For $n$ time points, $r$ reactions and $m$ metabolites, $A$ is an $(n \times m \times n \times r)$-matrix and $b$ an $(n \cdot m \times 1)$-vector. A linear optimization problem of the following form is solved:

\[
\begin{align*}
\text{minimize or maximize } & \quad F \\
\text{subject to } & \quad A \cdot v_t = b \\
& \quad v_{\text{min}} \leq v_t \leq v_{\text{max}}
\end{align*}
\]
where $F$ is the objective function; $v_t = (v(t_1), \cdots, v(t_n))^T$ is the $(n \times 1)$-vector of reaction rates in the lumped model at each of the $n$ time points; $v_{\text{min}}$ and $v_{\text{max}}$ are vectors of lower and upper bounds for the reaction rates respectively. An optimum $F_{\text{opt}}$ for the objective function and optimal reaction rate profiles are obtained.

(D)FBA assumes that cells have evolved towards optimal cellular behavior for survival. This means that for good performance of (D)FBA, it is crucial to have a suitable objective function for the given conditions [175].

In DFBA with Dynamic Optimization Approach (DOA) [120], an objective function has the form $\int_{t_0}^{t_f} f(t) \, dt$ where $t_0$ and $t_f$ are the start (steady state) and the end time of the experiment respectively. In previous studies [116, 120], the integral is approximated with the finite collocation method. Here we approximate the integral in the objective function by using the trapezoidal rule to reduce computational complexity. For equidistant time points $t_0, t_1, \cdots, t_n = t_f$:

$$\int_{t_0}^{t_f} f(t) \, dt = \sum_{i=1}^{n} \frac{1}{2} (f(t_i) + f(t_{i-1})) \cdot \Delta t$$

where $\Delta t = t_i - t_{i-1}$ for $i = 1, \cdots, n$. The objective function can be simplified by removing the factor $\frac{1}{2} \cdot \Delta t$ and the constant term $f(t_0)$ (sum of steady state fluxes). A linear optimization problem of the following form is solved:

minimize or maximize $F = f(t_1) + \sum_{i=2}^{n} (f(t_i) + f(t_{i-1}))$

subject to

$A \cdot v_t = b$

$v_{\text{min}} \leq v_t \leq v_{\text{max}}$

5.2.1.6 Dynamic FVA

To study the variability of the optimal reaction rate profiles leading to the same optimal value $F_{\text{opt}}$, FVA is performed. The following optimiza-
Integrating time-resolved metabolomics data into dynamic flux balance analysis

Optimization problems are solved:

\[
\begin{align*}
\text{minimize} & \quad v_j(t_i) \quad j = 1, \ldots, r \ ; \ i = 1, \ldots, n \\
\text{subject to} & \quad A \cdot v_t = b \\
& \quad v_{\text{min}} \leq v_t \leq v_{\text{max}} \\
& \quad F = F_{\text{opt}}
\end{align*}
\]

\[
\begin{align*}
\text{maximize} & \quad v_j(t_i) \quad j = 1, \ldots, r \ ; \ i = 1, \ldots, n \\
\text{subject to} & \quad A \cdot v_t = b \\
& \quad v_{\text{min}} \leq v_t \leq v_{\text{max}} \\
& \quad F = F_{\text{opt}}
\end{align*}
\]

Lower and upper bounds on the optimal solutions for the reaction rates at each time point are obtained.

\subsection{5.2.1.7 Software for optimization and smoothing}

Optimization problems were solved using the CPLEX [193] for MATLAB Toolbox. Smoothing was performed using the splinefit code [115] for MATLAB [128].
5.2.2 Case study: response of *S. cerevisiae* to a glucose pulse

5.2.2.1 Data set

The proposed method was applied to time-resolved metabolomics data of the central carbon metabolism in *S. cerevisiae*. Cells were cultivated in aerobic chemostats (dilution rate $D = 0.1 \ h^{-1}$), as described elsewhere [25, 67, 81]. Short-term perturbation-response experiments were carried out by introducing a sudden increase of 10 mM in the extracellular glucose concentration (also called a 10 mM glucose pulse). The data set consists of time series of quantitative data on extracellular and intracellular metabolite concentration levels (in mM). Extracellular metabolite levels were obtained enzymatically as described elsewhere [25]. Each time series consists of 14-16 time points. Table 5.1 gives an overview of the extracellular metabolites in the data set. Intracellular metabolite concentrations were determined by LC-MS/MS as described elsewhere [25]. The time series consist of 11-15 time points. The metabolites measured are from glycolysis and its branches, the tricarboxylic acid (TCA) cycle and the pentose phosphate pathway. Table 5.2 gives an overview. Data were measured on the second scale. The shortest window of observation was from 0 to 130 seconds. The data are plotted in the Supplementary Material.
Integrating time-resolved metabolomics data into dynamic flux balance analysis

Table 5.1: Extracellular metabolites in the data set.

<table>
<thead>
<tr>
<th>metabolite</th>
<th>number of biological replicates</th>
</tr>
</thead>
<tbody>
<tr>
<td>glucose (GLUC)</td>
<td>4</td>
</tr>
<tr>
<td>ethanol (ETH)</td>
<td>4</td>
</tr>
<tr>
<td>acetate (ACE)</td>
<td>4</td>
</tr>
<tr>
<td>glycerol (GLYCE)</td>
<td>3</td>
</tr>
<tr>
<td>acetaldehyde (ACALD)</td>
<td>2</td>
</tr>
</tbody>
</table>

Table 5.2: Intracellular metabolites in the data set.

<table>
<thead>
<tr>
<th>metabolite</th>
<th>pathway/branch</th>
<th>number of biological replicates</th>
</tr>
</thead>
<tbody>
<tr>
<td>glucose-6-phosphate (G6P)</td>
<td>glycolysis</td>
<td>4</td>
</tr>
<tr>
<td>fructose-6-phosphate (F6P)</td>
<td>glycolysis</td>
<td>4</td>
</tr>
<tr>
<td>fructose-1,6-bisphosphate (FBP)</td>
<td>glycolysis</td>
<td>4</td>
</tr>
<tr>
<td>3-phosphoglycerate (3PG)</td>
<td>glycolysis</td>
<td>4</td>
</tr>
<tr>
<td>phosphoenolpyruvate (PEP)</td>
<td>glycolysis</td>
<td>4</td>
</tr>
<tr>
<td>pyruvate (PYR)</td>
<td>glycolysis</td>
<td>4</td>
</tr>
<tr>
<td>glyceraldehyde-3-phosphate (GAP)</td>
<td>glycolysis</td>
<td>1</td>
</tr>
<tr>
<td>dihydroxyacetonephosphate (DHAP)</td>
<td>glycolysis</td>
<td>1</td>
</tr>
<tr>
<td>fructose-2,6-bisphosphate (F26bP)</td>
<td>regulator of glycolysis</td>
<td>4</td>
</tr>
<tr>
<td>trehalose-6-phosphate (T6P)</td>
<td>trehalose branch</td>
<td>4</td>
</tr>
<tr>
<td>glyceral-3-phosphate (G3P)</td>
<td>glycerol branch</td>
<td>4</td>
</tr>
<tr>
<td>glucose-1-phosphate (G1P)</td>
<td>glycogen branch</td>
<td>4</td>
</tr>
<tr>
<td>mannose-6-phosphate (M6P)</td>
<td>mannose branch</td>
<td>4</td>
</tr>
<tr>
<td>citrate (CIT)</td>
<td>TCA cycle</td>
<td>4</td>
</tr>
<tr>
<td>oxoglutarate (OGL)</td>
<td>TCA cycle</td>
<td>4</td>
</tr>
<tr>
<td>succinate (SUC)</td>
<td>TCA cycle</td>
<td>4</td>
</tr>
<tr>
<td>fumarate (FUM)</td>
<td>TCA cycle</td>
<td>4</td>
</tr>
<tr>
<td>malate (MAL)</td>
<td>TCA cycle</td>
<td>4</td>
</tr>
<tr>
<td>glyoxylic acid (GLYOX)</td>
<td>glyoxylic shunt</td>
<td>4</td>
</tr>
<tr>
<td>6-phosphoglucuronate (6PG)</td>
<td>pentose phosphate pathway</td>
<td>4</td>
</tr>
<tr>
<td>erythrose-4-phosphate (E4P)</td>
<td>pentose phosphate pathway</td>
<td>1</td>
</tr>
<tr>
<td>ribose-5-phosphate (R5P)</td>
<td>pentose phosphate pathway</td>
<td>1</td>
</tr>
<tr>
<td>xylose-5-phosphate (X5P)</td>
<td>pentose phosphate pathway</td>
<td>1</td>
</tr>
<tr>
<td>ribulose-5-phosphate (Rbu5P)</td>
<td>pentose phosphate pathway</td>
<td>1</td>
</tr>
<tr>
<td>sedoheptulose-7-phosphate (S7P)</td>
<td>pentose phosphate pathway</td>
<td>1</td>
</tr>
<tr>
<td>adenosine monophosphate (AMP)</td>
<td>cofactors</td>
<td>2</td>
</tr>
<tr>
<td>adenosine diphosphate (ADP)</td>
<td>cofactors</td>
<td>2</td>
</tr>
<tr>
<td>adenosine triphosphate (ATP)</td>
<td>cofactors</td>
<td>2</td>
</tr>
</tbody>
</table>
5.2.2.2 Steady state fluxes

Flux balance analysis [147] is applied on the most recent genome scale model of *S. cerevisiae* [65]. Fluxes of reactions supposed not to occur under the given experimental conditions (e.g. because a certain substrate is not in the medium), are equated to zero. The lower bound of the growth rate is set to 0.1 $h^{-1}$ (dilution rate, see 5.2.2.1.). The lower and upper bounds of the remaining reaction rates are taken from the genome scale model [65]. A list of the reactions in the genome scale model, together with the lower and upper bounds applied under the experimental conditions described in 5.2.2.1., is given in the Supplementary Material. When glucose is limited, it is realistic to suppose that the glucose uptake is minimized. The following optimization problem was solved:

$$\text{minimize } Z = \text{glucose uptake rate}$$

subject to

$$S \cdot v = 0 \text{ (steady state mass balances)}$$

for each reaction rate $v_j (j = 1, \cdots, r)$: $v_{j,\text{min}} \leq v_j \leq v_{j,\text{max}}$

where $S$ is the stoichiometric matrix of the genome scale model, $v$ is the flux vector and $r$ is the number of reactions. The optimization above results in optimal steady state fluxes and an optimal value (minimum) $Z_{\text{opt}}$ for the glucose uptake rate. Lower and upper bounds for the optimal steady state fluxes are obtained with FVA.

5.2.2.3 Calculation of derivatives

The data were smoothed with B-splines (piecewise polynomials) [64] and the time derivatives of the spline functions were calculated. The derivatives of the splines are evaluated at $t = 10, 20, 30, \cdots, 130$ seconds to obtain 13 equidistant time points which are within the window of observation of all time series in the data set. When a metabolite is measured multiple times, the average of the derivatives is calculated. A $(33 \times 13)$-matrix $dC/dt$ (33 metabolites, 13 time points) of time derivatives of the metabolite concentrations $C$ is obtained.
5.2.2.4 Lumped model

After lumping and removing reactions that do not occur under the given experimental conditions, the model consists of 33 metabolites and 62 reactions. The reaction scheme and a detailed list of the reactions are given in the Supplementary Material.

5.2.2.5 DFBA: Objective functions

In this study, seven objective functions are compared on their ability to give biologically meaningful estimates of the reaction rate profiles. Table 5.3 gives an overview of the seven objective functions. The objective functions max $v_{\text{biomass}}$, max $v_{\text{ATP}}$, max $v_{\text{ATP cyt}}$, max glucose uptake and max $v_{\text{ethanol}}$ are of the form $\int_{t_0}^{t_f} f(t) \, dt$, where $f(t) = \sum k_i \cdot v_i(t)$ is a linear combination of the reaction rates, and can be approximated with the trapezoidal rule, like described in 5.2.1.5. Each of the seven objective functions is further explained in the next subsections.

5.2.2.5.1 Maximize biomass yield

In most (D)FBA studies of microorganisms, it is assumed that survival is equivalent with growth [46] and the objective is maximization of the biomass (max $v_{\text{biomass}}$). However, there are conditions where the cell does not grow optimally [175].

5.2.2.5.2 Maximize ATP yield

The objective of maximizing ATP yield (max $v_{\text{ATP}}$) is based on the assumption that cells maximize energy production when oxygen is available [156].

5.2.2.5.3 Maximize ATP yield in the cytosol

Studies of perturbation-response experiments showed that after a glucose pulse, there is a switch from respiratory to respiro-fermentative metabolism and as a consequence low TCA activity [56]. This means that there is low ATP production in the mitochondria. Therefore maximizing only cytosolic ATP (max $v_{\text{ATP cyt}}$) was also studied.
Integrating time-resolved metabolomics data into dynamic flux balance analysis

Table 5.3: Overview of the objective functions used in this study.

<table>
<thead>
<tr>
<th>objective function</th>
<th>explanation</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\max v_{biomass}$</td>
<td>maximal biomass yield</td>
</tr>
<tr>
<td>subject to $A \cdot v_t = b$</td>
<td>most common objective function in FBA studies [109, 147]</td>
</tr>
<tr>
<td>irreversibility constraints (see FBA)</td>
<td></td>
</tr>
</tbody>
</table>

| max $v_{ATP}$ | maximal energetic efficiency [175, 109] |
| subject to $A \cdot v_t = b$ | |
| irreversibility constraints | |
| rate to TCA cycle $\leq$ steady state flux to TCA cycle | switch to respiro-fermentative metabolism |
| growth rate $= 0.1 \text{ h}^{-1}$ | (excess glucose fermented to ethanol)[104] |
| growth rate does not change significantly during a small time interval | |

| max $v_{ATP_{cyt}}$ | maximize energetic efficiency in the cytosol |
| subject to $A \cdot v_t = b$ | respiro-fermentative metabolism [56] |
| irreversibility constraints | $\rightarrow$ low TCA cycle activity |
| rate to TCA cycle $\leq$ steady state flux to TCA cycle | $\rightarrow$ low ATP production in mitochondria |
| growth rate $= 0.1 \text{ h}^{-1}$ | excess glucose fermented to ethanol |
| growth rate does not change significantly during a small time interval | |

| max glucose uptake rate | organism consumes the excess of glucose as fast as possible |
| subject to $A \cdot v_t = b$ | |
| irreversibility constraints | |
| rate to TCA cycle $\leq$ steady state flux to TCA cycle | excess glucose fermented to ethanol |
| growth rate $= 0.1 \text{ h}^{-1}$ | growth rate does not change significantly during a small time interval |

| max $v_{ethanol}$ | maximize ethanol yield |
| subject to $A \cdot v_t = b$ | excess glucose fermented to ethanol |
| irreversibility constraints | growth rate does not change significantly during a small time interval |
| growth rate $= 0.1 \text{ h}^{-1}$ | |

| min sum of absolute fluxes | minimize enzyme usage |
| subject to $A \cdot v_t = b$ | same constraints as FBA |
| irreversibility constraints | |
| $0.1 \leq$ growth rate $\leq 1000$ | |

| min $\sum_{i=1}^{n} \sum_{j=1}^{l} (v_j (t_0) - v_j (t_i))^2$ | minimization of adjustment |
| subject to $A \cdot v_t = b$ | |
| irreversibility constraints | |
Integrating time-resolved metabolomics data into dynamic flux balance analysis

5.2.2.5.4 Maximize glucose uptake Maximizing glucose uptake as objective function is motivated by the fact that after a glucose pulse, *S. cerevisiae* takes up the excess of glucose fast [58].

5.2.2.5.5 Maximize ethanol yield *S. cerevisiae* is known to ferment an excess of glucose to ethanol [180]. The objective of maximizing ethanol (max $v_{ethanol}$) is therefore also included in this study.

5.2.2.5.6 Minimization of the overall flux Cells are assumed to minimize their enzyme usage, which corresponds to minimizing their overall flux [175, 70]. This can be mathematically formulated as minimizing the sum of the absolute values of all reaction rates [180]. The following dynamic optimization problem has to be solved:

$$\begin{align*}
\text{minimize} & \quad \int_{t_0}^{t_f} \left( \sum_{j=1}^{r} |v_j(t)| \right) \, dt \\
\text{subject to} & \\
A \cdot v_t & = b \\
v_{\text{min}} \leq v_t \leq v_{\text{max}}
\end{align*}$$

This problem can be converted to a linear optimization problem [114] by introducing $r$ more (slack) variables $v_{j+r} = |v_j| \ (j = 1, \cdots, r)$. The optimization problem above is equivalent with [52]

$$\begin{align*}
\text{minimize} & \quad F = \int_{t_0}^{t_f} \left( \sum_{j=1}^{r} v_{j+r} \right) \, dt \\
\text{subject to} & \\
A \cdot v_t & = b \\
v_{\text{min}} \leq v_t \leq v_{\text{max}} \\
v_{j+r} & \geq -v_j \ (j = 1, \cdots, r) \\
v_{j+r} & \geq v_j \ (j = 1, \cdots, r)
\end{align*}$$
Approximating the integral in the objective function by using the trapezoidal rule (see 5.2.1.5) results in a linear objective function.

5.2.2.5.7 Minimization of Metabolic Adjustment (MOMA) Living organisms are assumed to adjust their metabolism with minimal effort after a perturbation. This is called Minimization of Metabolic Adjustment (MOMA) [178] and can be formulated as minimizing the sum of the squared deviations from the steady state fluxes. The following quadratic optimization problem is solved:

\[
\minimize F = \sum_{i=1}^{n} \sum_{j=1}^{r} \left( v_j (t_0) - v_j (t_i) \right)^2
\]

subject to
\[
A \cdot v_t = b
\]
\[
v_{min} \leq v_t \leq v_{max}
\]

This results in an optimum \( F_{opt} \) for the objective function and optimal dynamic rate profiles.

5.2.2.6 Combining objective functions

It has been proposed in the literature that flux (reaction rate) changes after a perturbation occur as a result of a combination of maximal metabolic or energetic efficiency under the given conditions, and minimization of adjustment [176, 132] or minimization of the sum of absolute fluxes [180].

Previous studies [136, 176] showed that objective functions compete against each other. This means that one objective function can only be improved if another is worsened. Optimal solutions for competing objectives are called Pareto optimal [176]. The set of Pareto optimal solutions is called the Pareto front [89]. A method to calculate Pareto optima is optimizing a weighted sum of the objectives [219]. The follo-
Integrating time-resolved metabolomics data into dynamic flux balance analysis

wing multi-objective optimization problem is solved:

\[
\min F = F_1 - w \cdot F_2 \\
\text{subject to} \quad A \cdot v_t = b \\
v_{\text{min}} \leq v_t \leq v_{\text{max}}
\]

where \(F_1\) is MOMA or the sum of absolute fluxes and \(F_2\) is one of the other objective functions of Table 5.3. An overview of the multi-objective optimization problems solved in this study is given in Table 5.4.

Table 5.4: Overview of the multi-objective optimization problems solved in this study.

<table>
<thead>
<tr>
<th>objective function</th>
<th>equation</th>
</tr>
</thead>
<tbody>
<tr>
<td>(\min \left{ \sum_{i=1}^{n} \sum_{j=1}^{r} (v_j(t_0) - v_j(t_i))^2 - w \cdot v_{\text{ATP}} \right} )</td>
<td>(\sum_{i=1}^{n} \sum_{j=1}^{r} (v_j(t_0) - v_j(t_i))^2 - w \cdot v_{\text{ATP}} )</td>
</tr>
<tr>
<td>(\min \left{ \sum_{i=1}^{n} \sum_{j=1}^{r} (v_j(t_0) - v_j(t_i))^2 - w \cdot v_{\text{gluc uptake}} \right} )</td>
<td>(\sum_{i=1}^{n} \sum_{j=1}^{r} (v_j(t_0) - v_j(t_i))^2 - w \cdot v_{\text{gluc uptake}} )</td>
</tr>
<tr>
<td>(\min \left{ \sum_{i=1}^{n} \sum_{j=1}^{r} (v_j(t_0) - v_j(t_i))^2 - w \cdot v_{\text{ETH}} \right} )</td>
<td>(\sum_{i=1}^{n} \sum_{j=1}^{r} (v_j(t_0) - v_j(t_i))^2 - w \cdot v_{\text{ETH}} )</td>
</tr>
<tr>
<td>(\min \left{ \sum_{i=1}^{n} \sum_{j=1}^{r} (v_j(t_0) - v_j(t_i))^2 - w \cdot v_{\text{biomass}} \right} )</td>
<td>(\sum_{i=1}^{n} \sum_{j=1}^{r} (v_j(t_0) - v_j(t_i))^2 - w \cdot v_{\text{biomass}} )</td>
</tr>
<tr>
<td>(\min \left{ \sum_{i=1}^{n} \sum_{j=1}^{r} (v_j(t_0) - v_j(t_i))^2 - w \cdot v_{\text{ATP cyt}} \right} )</td>
<td>(\sum_{i=1}^{n} \sum_{j=1}^{r} (v_j(t_0) - v_j(t_i))^2 - w \cdot v_{\text{ATP cyt}} )</td>
</tr>
<tr>
<td>(\min \left{ \text{sum of absolute fluxes} - w \cdot v_{\text{ATP}} \right} )</td>
<td>(\text{sum of absolute fluxes} - w \cdot v_{\text{ATP}} )</td>
</tr>
<tr>
<td>(\min \left{ \text{sum of absolute fluxes} - w \cdot v_{\text{gluc uptake}} \right} )</td>
<td>(\text{sum of absolute fluxes} - w \cdot v_{\text{gluc uptake}} )</td>
</tr>
<tr>
<td>(\min \left{ \text{sum of absolute fluxes} - w \cdot v_{\text{ETH}} \right} )</td>
<td>(\text{sum of absolute fluxes} - w \cdot v_{\text{ETH}} )</td>
</tr>
<tr>
<td>(\min \left{ \text{sum of absolute fluxes} - w \cdot v_{\text{biomass}} \right} )</td>
<td>(\text{sum of absolute fluxes} - w \cdot v_{\text{biomass}} )</td>
</tr>
<tr>
<td>(\min \left{ \text{sum of absolute fluxes} - w \cdot v_{\text{ATP cyt}} \right} )</td>
<td>(\text{sum of absolute fluxes} - w \cdot v_{\text{ATP cyt}} )</td>
</tr>
</tbody>
</table>

Each value of \(w\) determines a point on the Pareto front [89] (see Figure 5.3). In this study, the results of the optimization are shown for \(w = 10^p\) \((p = -3, \cdots, 6)\). Making \(w\) larger or smaller did not change the result anymore.
Integrating time-resolved metabolomics data into dynamic flux balance analysis

Figure 5.3: Different Pareto optimal solutions $P_1, P_2, P_3, \ldots$ can be calculated by varying the weight $w$ in the multi-objective optimization.

5.2.2.7 Properties of biologically meaningful solutions

The solutions of the DFBA (+ DFVA) are checked for the following properties:

1. The interval of optimal solutions is small.

2. The directionality of the network corresponds with literature [23].

3. Switch to respiro-fermentative conditions [104]. Low TCA cycle activity under respiro-fermentative conditions [56].

4. The reaction rates are lower for upper glycolysis (G6P $\rightarrow$ F6P $\rightarrow$ FBP $\rightarrow$ GAP + DHAP) than for lower glycolysis (GAP $\rightarrow$ 3PG $\rightarrow$ PEP $\rightarrow$ PYR) [56].

5. The qualitative behavior of the reaction rates corresponds with the model simulations of Vaseghi et al., where all reaction rates of the
pentose phosphate pathway show a fast increase immediately after the pulse, followed by a very slow decrease [215].

Properties 2, 3 and 4 are based on physiology. Objective functions can be rejected if the reaction rate profiles do not correspond to one or more of the properties 2, 3 and 4. Property 5 is based on a kinetic model. Contradictions between a kinetic model and a DFBA model mean that one of the two models (or both) does not describe the reaction rates well.
5.3 Results

5.3.1 Optimizing a single objective function

Table 5.5 shows the results of checking the five properties described in 5.2.2.7. for each of the seven optimization problems shown in Table 5.3.

<table>
<thead>
<tr>
<th>objective function</th>
<th>order of the $F_{opt}$</th>
<th>contradiction with</th>
</tr>
</thead>
<tbody>
<tr>
<td>max $v_{\text{biomass}}$</td>
<td>$10^4$</td>
<td>1</td>
</tr>
<tr>
<td>max $v_{\text{ATP}}$</td>
<td>$10^5$</td>
<td>1</td>
</tr>
<tr>
<td>max $v_{\text{ATP}_{\text{cyt}}}$</td>
<td>$10^{-1}$</td>
<td>1</td>
</tr>
<tr>
<td>max $v_{\text{glucose uptake rate}}$</td>
<td>$10^5$</td>
<td>1</td>
</tr>
<tr>
<td>max $v_{\text{ethanol}}$</td>
<td>$10^3$</td>
<td>1</td>
</tr>
<tr>
<td>min sum of absolute fluxes</td>
<td>$10^3$</td>
<td>2 and 5</td>
</tr>
<tr>
<td>MOMA</td>
<td>$10^2$</td>
<td>5</td>
</tr>
</tbody>
</table>

$F_{opt} = \text{optimum for the objective function}$. With max $v_{\text{biomass}}$ we mean max $\int_0^t v_{\text{biomass}}(t) \, dt$. Idem for max $v_{\text{ATP}}$, max $v_{\text{ATP}_{\text{cyt}}}$, max $v_{\text{glucose uptake rate}}$, max $v_{\text{ethanol}}$.

Under steady state, the reaction rates for the TCA cycle in the FBA solution are in the order of $10^{-1} - 10^0$. A TCA cycle activity of higher order than the steady state TCA cycle activity is in contradiction with physiology [56].

Maximizing $v_{\text{biomass}}$, $v_{\text{ATP}}$, $v_{\text{ATP}_{\text{cyt}}}$, glucose uptake or $v_{\text{ethanol}}$ resulted in very large variability among the optimal solutions (large difference between lower and upper bound resulting from DFVA, see Figure 5.4 for an example). The solution space is not enough constrained to draw conclusions about the reaction rate profiles under the experimental conditions.

114
Integrating time-resolved metabolomics data into dynamic flux balance analysis

Figure 5.4: DFBA (red asterisks) and DFVA (blue circles) solution for six reaction rates when ATP in the cytosol is maximized. Similar profiles for the other reaction rates. Idem for max $v_{\text{biomass}}$, max $v_{\text{ATP}}$, max glucose uptake and max $v_{\text{ethanol}}$. Abbreviations: PYR, pyruvate; ATP, adenosine triphosphate; ADP, adenosine diphosphate; AMP, adenosine monophosphate; Rbu5P, ribulose-5-phosphate; E4P, erythrose-4-phosphate; PEP, phosphoenolpyruvate.

Minimization of the sum of absolute fluxes resulted in small variability among the optimal solutions (see Figure 5.5).
Integrating time-resolved metabolomics data into dynamic flux balance analysis

Figure 5.5: DFBA (red asterisks) and DFVA (blue circles) solution for six reaction rates when only minimizing the sum of absolute fluxes. Abbreviations, see Figure 5.4.

However, the directionality (sign of the reaction rates) of glycolysis did not correspond to the literature [23] (see Figure 5.6).
The sign of the rates of the reactions of the TCA cycle and the pentose phosphate pathway was in accordance with the literature [56]. The predictions of the DFBA showed low TCA cycle activity (reaction rates in the order of $10^{-1}$ $gdW^{-1}h^{-1}$), which corresponds to previous studies. The qualitative behavior of the reaction rates of the pentose phosphate pathway is not in accordance with the model simulations of Vaseghi et al. [215]. There, all reaction rates show a fast increase followed by a very slow decrease, while for the DFBA predictions the reaction rates decrease rapidly (see Figure 5.7).
Figure 5.7: Reaction rate profile for one of the reactions of the pentose phosphate pathway (Ru5PE). (a) DFBA (red asterisks) and DFVA (blue circles) solution when minimizing the sum of the absolute fluxes. (b) Reaction rate profile for the model of Vaseghi et al. [215], figure adapted from Vaseghi et al. The other reaction rates of the pentose phosphate pathway show similar profiles. Ru5PE: ribulose phosphate epimerase.

The results of MOMA optimization show small variability among the optimal solutions (see Figure 5.8).
Integrating time-resolved metabolomics data into dynamic flux balance analysis

The directionality of the network (sign of the reaction rates) corresponds to the literature [23]. The reaction rates for the TCA cycle are low (in the order of $10^{-1} - 10^0 \text{ gDW}^{-1} \text{ h}^{-1}$), which is in accordance with previous research [56]. The reaction rate profiles of the pentose phosphate pathway show a large increase after the glucose pulse, followed by a fast decrease, which does not correspond to the model simulations of Vaseghi et al. [215](see Figure 5.9).

Figure 5.8: DFBA (red asterisks) and DFVA (blue circles) solution for six reaction rates when only optimizing MOMA. Abbreviations, see Figure 5.4.
Integrating time-resolved metabolomics data into dynamic flux balance analysis

Figure 5.9: Reaction rate profile for one of the reactions of the pentose phosphate pathway (Ru5PE). (a) DFBA-MOMA (red asterisks) and DFVA-MOMA (blue circles) solution. (b) Reaction rate profile for the model of Vaseghi et al. [215], figure adapted from Vaseghi et al. The other reaction rates of the pentose phosphate pathway show similar profiles. Ru5PE: ribulose phosphate epimerase.

5.3.2 Multi-objective optimization

In Table 5.6, the results of checking the five properties in paragraph 2.2.7. for the solutions of the multi-objective optimization problems of Table 5.4 are shown.
Integrating time-resolved metabolomics data into dynamic flux balance analysis

Figures illustrating the contradictions with the literature mentioned in Table 5.6 are shown in the Supplementary Material. The DFBA model satisfies all of the five properties when MOMA is combined with maximization of ethanol production and the weight is in the range of $10^4 - 10^6$ (the orders of $F_1$, $F_2$ and $w \cdot F_2$ are $10^7$, $10^5$ and $10^9 - 10^{11}$ respectively). This means that maximizing ethanol production ($F_2$) is more important than MOMA ($F_1$). There is no significant difference among the solutions for $w = 10^4$, $10^5$ and $10^6$ (see Supplementary Material).

5.4 Discussion

Studying cellular dynamics is important for understanding adaptation mechanisms of living organisms. Kinetic models can help to explore dynamic properties of the cell. However, building kinetic models is hampered by incomplete kinetic information. Therefore, there is a need for mathematical methods that require only a minimal amount of kinetic information. DFBA does not need detailed kinetics, but is limited to smaller systems or quasi steady state because of its mathematical complexity and the large number of unknowns.

In this study, we presented a mathematical framework to integrate time-resolved metabolomics data into DFBA. Combining DFBA with time-resolved metabolomics data reduces both mathematical complexity and the number of unknowns.

The new method described in this chapter can be applied in two ways. When the biological goal of the system is unknown, one can test hypotheses (objective functions). For a known objective function, DFBA combined with time-resolved metabolomics data can be used to calculate reaction rate profiles.

In this chapter, a case study of the first application, testing objective functions, was presented. DFBA was combined with data of short term perturbation-response experiments (10 mM glucose pulse) of \textit{S. cerevisiae} to test plausible objective functions. Both single objective and multi-objective optimization were considered.

Minimizing the overall flux or MOMA resulted in a small solution space.
Integrating time-resolved metabolomics data into dynamic flux balance analysis

**Objective**

\[ F \text{order} = F \text{order} \]

Contradiction with literature?

\[ \min \left\{ \left( \sum_{i} \sum_{j} \left( v_{j}^{(t)} - v_{j}^{(t)} \right) \right)^{2} \right\} - w \cdot v_{ATP} \]

\[ 0.001 \times 10^{-3} \times 10^{-5} \times 10^{-7} \times 10^{-9} \times 10^{-11} \]

Contradiction with 3 and 5

\[ \min \left\{ \left( \sum_{i} \sum_{j} \left( v_{j}^{(t)} - v_{j}^{(t)} \right) \right)^{2} \right\} - w \cdot v_{ETH} \]

\[ 0.001 \times 10^{-3} \times 10^{-5} \times 10^{-7} \times 10^{-9} \times 10^{-11} \]

Contradiction with 3

\[ \min \left\{ \left( \sum_{i} \sum_{j} \left( v_{j}^{(t)} - v_{j}^{(t)} \right) \right)^{2} \right\} - w \cdot v_{biomass} \]

\[ 0.001 \times 10^{-3} \times 10^{-5} \times 10^{-7} \times 10^{-9} \times 10^{-11} \]

Contradiction with 2

\[ \min \left\{ \left( \sum_{i} \sum_{j} \left( v_{j}^{(t)} - v_{j}^{(t)} \right) \right)^{2} \right\} - w \cdot v_{ATP cyt} \]

\[ 0.001 \times 10^{-3} \times 10^{-5} \times 10^{-7} \times 10^{-9} \times 10^{-11} \]

Contradiction with 4 and 5

**Table 5.6: Results of checking the five properties described in 5.2.2.7. when solving the multi-objective optimization problems of Table 5.4.**

<table>
<thead>
<tr>
<th>Property</th>
<th>Contradiction with</th>
<th>Order</th>
<th>Property</th>
<th>Contradiction with</th>
<th>Order</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>1</td>
<td>3</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>1</td>
<td>4</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td>1</td>
<td>1</td>
<td>5</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>
Integrating time-resolved metabolomics data into dynamic flux balance analysis

Maximizing one of the other five objectives of Table 5.3 (max $v_{\text{biomass}}$, max $v_{\text{ATP}}$, max $v_{\text{ATP}_{\text{cyt}}}$, max glucose uptake and max $v_{\text{ethanol}}$) ended up in a solution space that was too large to draw any conclusions about the reaction rate profiles. Probably this is due to the fact that minimizing the overall flux and MOMA constrain all the reaction rates, while the other objective functions only constrain the reaction rate(s) from or to one specific metabolite.

When minimizing the overall flux, the directionality of the network was not in accordance with the physiological information in the literature [56]. This means that the pathways followed by *S. cerevisiae* after a glucose pulse are not the result of minimizing enzyme usage only.

The MOMA solution was in accordance with the physiological information in the literature, but the qualitative behavior did not correspond to the model simulations of Vaseghi *et al.* [215]. Differences in reaction rate profiles between the DFBA model and the kinetic model of Vaseghi *et al.* can be due to different factors:

- difference in strain (see Table 5.7);
- differences in experimental conditions (see Table 5.7);
- only minimizing metabolic adjustment is not the biological goal of *S. cerevisiae* under the conditions in this study.

Further research is needed to conclude which factor(s) lead to the difference between the DFBA model and the kinetic model.

Table 5.7: Strain and experimental conditions in this study and in the study of Vaseghi *et al.* [215].

<table>
<thead>
<tr>
<th>This study</th>
<th>Vaseghi <em>et al.</em> [215]</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>S. cerevisiae</em> CEN.PK 113-7D</td>
<td><em>S. cerevisiae</em> CBS 7336</td>
</tr>
<tr>
<td>dilution rate = 0.1 h$^{-1}$</td>
<td>dilution rate = 0.1 h$^{-1}$</td>
</tr>
<tr>
<td>temperature = 30°C</td>
<td>temperature = 30°C</td>
</tr>
<tr>
<td>pH = 5</td>
<td>pH = 5</td>
</tr>
<tr>
<td>steady state biomass = 3.75 gDW l$^{-1}$</td>
<td>steady state biomass = 15 gDW l$^{-1}$</td>
</tr>
<tr>
<td>pulse = 10 mM glucose</td>
<td>pulse = 5.6 mM glucose</td>
</tr>
</tbody>
</table>

The following combinations of objective functions can be rejected based on physiology (contradiction with 2, 3 or 4, see Table 5.6):
• maximize $v_{ATP}$, $v_{glucose\ uptake}$ or $v_{biomass}$ as main objective and MOMA as second objective;

• minimize overall flux and maximize $v_{ATP}$, $v_{ETH}$ or $v_{ATP_{cyt}}$ (for all weights);

• minimize overall flux as main objective, maximize $v_{glucose\ uptake}$ or $v_{biomass}$ as second objective;

• minimize overall flux and maximize $v_{biomass}$ equally important;

• minimize overall flux and maximize $v_{biomass}$ ($w \cdot F2 \geq 100 \cdot F1$).

Only one combination of objectives also corresponds with the qualitative behavior of the model simulations of Vaseghi et al. [215], namely maximize $v_{ETH}$ as the main objective and MOMA as the second objective ($w \cdot F2 \geq 100 \cdot F1$). In the combined objective function, the term $v_{ETH}$ remained of the same order as for optimizing only $v_{ETH}$ (order $10^5$). The $v_{ETH}$ with a value of $1.25 \cdot 10^5$ changed extremely little, only visible at the sixth decimal place (this is probably a non-significant change). The MOMA term changed from $5.97 \cdot 10^2$ to $3.64 \cdot 10^7$ from just only MOMA to MOMA combined with max $v_{ETH}$ (60000 times higher). These results show that maximizing ethanol yield is very important compared to MOMA. There is only a subtle change in the $v_{ETH}$ term if we compare only max $v_{ETH}$ with max $v_{ETH}$ combined with MOMA.

In cases where there is only a subtle change in the main objective when adding a second objective, the solution of the weighted optimization is almost the same as sequential optimization of the two objectives. In the case study, first maximizing $v_{ETH}$ and then minimizing metabolic adjustment given the maximum for $v_{ETH}$ ($1.25 \cdot 10^5$) resulted in a MOMA optimum of which differs not significantly with the MOMA term obtained by optimizing a weighted sum with $w = 10^4 - 10^6$.

In the literature, it is often claimed that S. cerevisiae has evolved to produce ethanol because of its toxicity for most other microorganisms and because ethanol can be consumed as carbon source when glucose is depleted [153]. However, there is no experimental evidence for this hypothesis [132].
The results of this study show that optimizing the yield or uptake of a single compound was insufficient to obtain a small range of reaction rate profiles with DFBA. A small solution space could only be obtained from at least two objectives or from one objective that is a function of all the reaction rates.

The method presented in this chapter has several advantages compared to standard DFBA. Incorporating time-resolved metabolomics data into DFBA converts the differential equation constraints of standard DFBA to linear constraints. In this way it reduces the complexity of the optimization problem and as a consequence also the computational time. Furthermore, including experimental data reduces the solution space.

A limitation of the method is that it needs quantitative time-resolved data of metabolite concentrations (in mM or mmol/gdW), while many current data sets consist of only steady state or semi-quantitative (peak areas) measurements. It is expected that there will be more data sets suitable for applying this method in the future because of continuous improvement of analytical methods.

When applying the DFBA combined with time-resolved metabolomics data for reaction rate estimation, preknowledge of the importance of each objective function is required in order to determine the weight of each objective function.

In this study, the method was validated using literature. The conclusions therefore depend on the reliability of previous studies. Properties 2, 3 and 4 (see paragraph 5.2.2.7.) are based on experimental data. Property 5 is based on simulations with a kinetic model. Care has to be taken with conclusions that are solely derived from simulations, because there can be differences in strain and experimental conditions between the experimental data and the kinetic model.

A more accurate validation of the method could be performed by comparing the estimated reaction rate profiles with reaction rates determined by dynamic $^{13}$C metabolic flux analysis. If such data become available, this would be a recommendation for the future.
5.5 Conclusion

Studying reaction rates and optimization principles is important to understand cellular functioning and adaptation to changing environments. In this study, DFBA was combined with experimental data to overcome some of the limitations of standard DFBA. The DFBA-framework can be used for testing hypotheses about optimal cellular behavior under perturbations. When it is known how an organism behaves optimal under a certain condition, the DFBA-framework can be used to estimate reaction rate profiles. The method does not require detailed kinetics but, when available, kinetic information can be easily incorporated. Kinetic models can be used to calculate reaction rates or estimate non-measured metabolite profiles. If the parameters in rate laws are uncertain, they can be estimated with experimental data and the rate laws can be used as constraints for the (D)FBA [51, 159].

The method can be extended for three or more objective functions by using a weighted sum of objective functions. When estimation of the weights is a problem, successive optimization can be performed as an alternative.

In this study, the method was validated by comparing the results with literature. If $^{13}$C labeling data are available, a more accurate validation method would be to compare the results with $^{13}$C dynamic time series data. If $^{13}$C data are only available for a limited number of reactions, they can be used to formulate extra constraints on the solution space of the DFBA.

Acknowledgments

This project was financed by the Netherlands Metabolomics Centre (NMC), which is part of the Netherlands Genomics Initiative - Netherlands Organization for Scientific Research.
Supplementary Data

The supplementary data is too extensive to be integrally included in this thesis. It will be made available online together with the manuscript that is currently in preparation.
Integrating time-resolved metabolomics data into dynamic flux balance analysis
Chapter 6

Conclusion and outlook

6.1 Conclusion

This project has investigated the possibilities to extract network properties from time-resolved metabolite concentration data. A study of the feasibility of reverse engineering of metabolic networks from time-resolved metabolomics data (Chapter 2) has shown that it is very difficult to estimate the structure of a metabolic network if the time constants in the network have different orders of magnitude. This means that if there are both fast and slow reactions in the network, the estimated connections are not reliable. Furthermore, the sampling frequencies required by network inference methods are not consistent with current measurement techniques. Also the noise levels in the data are too high for a good performance of network inference methods.

Chapter 3 focused on finding changes in pathways under different conditions. The feasibility of using Goeman’s global test, originally designed for gene-expression analysis, in metabolomics was studied. The results showed that Goeman’s global test is suitable to detect metabolic pathway differences between conditions. Goeman’s global test is able to detect very small changes between conditions. This is advantageous when all metabolites in a pathway have subtle changes that can not be
detected by testing single metabolites. The property of detecting very small changes can be disadvantageous when only a single metabolite in a pathway changes, because it can also lead to a significant result in this case.

Chapter 4 addressed correlation analysis in time-resolved metabolomics data under different conditions. From the results it can be concluded that correlations in time-resolved data can be used to detect differences in the distribution of reaction rates between conditions. When combining the results of the correlation analysis with network topology and directionality, a list of possible scenarios for regulation or redirection of reactions can be extracted. However, more information about the pathway is needed to select the correct regulation scenario from the list. A final study (chapter 5) focused on integration of time-resolved metabolomics data with dynamic flux balance analysis (DFBA). DFBA combined with time-resolved metabolomics data can be applied to test hypotheses about how an organism optimally adapts to a perturbation. Given an objective function, which represents the biological goal of an organism under a certain condition, the new DFBA method presented in chapter 5 can also be applied to estimate reaction rate profiles. The results of a case study on hypothesis testing showed that a small solution space could not be obtained by optimizing the yield or uptake of a single compound. At least two objectives or an objective that is a function of all reaction rates were needed. From all the points mentioned above, it can be concluded that time-resolved metabolic data provides a wealth of information about metabolic pathways that can not be derived from steady state studies.

6.2 Challenges for future research

6.2.1 Integrated network models

Metabolomics, transcriptomics, genomics and proteomics are mostly studied separately. In reality, there are interactions between the different layers of -omics data. Therefore, one of the challenges in systems biology is to infer integrated networks [200, 92]. This poses several difficulties.
Data from different platforms and different -omics technologies are very heterogeneous in terms of noise levels, linearity of the response and time scales [186]. Fast processes like signaling and metabolic reactions take place on a second or sub second scale, while slow processes like regulation and growth happen in minutes or hours [200]. Metrics based on mutual information can be used for heterogeneous data [93]. However, they are meant for static network analysis and do not account for dynamics. A challenge is to develop an association measure or metric that can take into account the huge differences in time-scale among different -omics levels [92]. Such a metric can then be used to calculate associations between time series of different -omics data and examine which additional information these associations provide compared to steady state associations (in a similar study as chapter 4). A second problem is that data are often incomplete, which causes gaps in the knowledge about the system. Methods have to be developed to fill these gaps [200]. Currently used methods for missing value imputation are, among others, replacement by the mean or the median of the rest of the samples, interpolation and non-linear PCA based methods [186]. Another challenge is the reduction of noise, which has a large impact on the performance of network inference (see chapter 2), especially on practical identifiability [196]. Noise can be reduced by measuring more replicates and taking the average.

6.2.2 Structural identifiability

Even if the data would be complete and noiseless, problems can still occur that are related to structural identifiability. This can hinder a complete kinetic description of the system because the parameters can not uniquely be identified [200]. When the network is not structurally identifiable, the model has to be reformulated in a way that the parameters are uniquely identifiable.

6.2.3 Differential networks

Metabolic network maps, like those in KEGG [82, 84, 83], give an overview of all metabolites and reactions that have been observed in an organism.
However, under a given condition, it rarely happens that all nodes and edges of the KEGG map are active [92]. Comparing the subgraphs of the network that are active across different conditions or between different time points can reveal information about regulation of the network. This is the goal of differential network analysis [72].

There are several challenges in studying subgraphs. The availability of data limits differential network analysis [92]. Because in most cases, not all metabolites are measured [37], only simplified (lumped) subgraphs can be inferred. Furthermore, calculating an association measure or test statistic requires enough biological replicates (e.g. \( \geq 10 \) replicates for correlations in steady state data)[21]. Moreover, missing data and measurement errors cause uncertainty in the results of inferring subgraphs [37]. Understanding of the experimental errors in the data is required to detect and eventually avoid false positives and false negatives [72]. Traditionally, measurement errors are quantified with the relative standard deviation (RSD). However, the RSD assumes a linear relationship between the mean and the standard deviation of the peak intensity. For GC-MS, LC-MS and NMR measurements, the standard deviation of the peak intensity is constant at low intensities and proportional to the mean at high intensities [160]. Van Batenburg et al. [206] showed that a variance model that includes both an additive (background noise) and a multiplicative parameter (additional source of measurement error occurring above a certain threshold for the mean peak intensity) is more appropriate to describe measurement errors in metabolomics data than the RSD.

### 6.2.4 Integration of modeling frameworks

Currently, a lot of computational systems biology frameworks exist, which have all their advantages and drawbacks (see Table 6.1). Recently, a few attempts have been made to combine those frameworks to overcome some of the drawbacks. Constrained-based models can have a large solution space, while in kinetic models often not all parameters are known. It has been shown that (incomplete) kinetic models can be used to reduce the solution space in constrained-based models [118, 51].
Conclusion and outlook

Another application of model integration is combining physiological models with mechanism-based systems biology. Mechanism-based systems biology models describe only a single compartment and do not address whole function of an organism. Physiological models do not include network information. By combining mechanism-based systems biology models with physiological models, one can overcome the limitations of both types of models. Combining physiology with mechanism-based models is challenging, because physiological processes take place on the slow time scale (minutes or hours), while reactions take place on the fast time scale (seconds or sub seconds) [200]. An example are SBPKPD models, which combine systems biology (SB) with pharmacokinetics-pharmacodynamics (PK-PD) models [195].

As a last example, we discuss the combination of flux balance analysis (FBA) and $^{13}$C metabolic flux analysis ($^{13}$C MFA). Both approaches have their limitations. $^{13}$C MFA does not take into account other well-known constraints than stoichiometry. Second, $^{13}$C MFA is influenced by propagation of the errors in the labeling experiments and external flux measurement [112]. Third, $^{13}$C MFA can only estimate fluxes related to carbon metabolism (because of the $^{13}$C labeling) [32]. Drawbacks of FBA are the dependence on the objective function [112] and that the optimal solution is often not unique [32]. Combining $^{13}$C MFA with FBA can have several advantages. The $^{13}$C MFA solution can be incorporated into FBA to determine fluxes not related to carbon metabolism [33]. Second, the validity of an objective function can be checked by comparing flux solutions of FBA and $^{13}$C MFA [38].

Studying how other combinations of current modeling frameworks can lead to a better description of the functioning of the cell is a challenge for future research.
Table 6.1: Overview of current computational systems biology frameworks, together with their advantages and drawbacks.

<table>
<thead>
<tr>
<th>framework</th>
<th>advantages</th>
<th>drawbacks</th>
<th>references</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bayesian networks</td>
<td>does not need detailed knowledge about the system</td>
<td>based on acyclic graphs while biological networks can contain cycles</td>
<td>[57, 80]</td>
</tr>
<tr>
<td></td>
<td>ability to handle uncertainty</td>
<td>determination of initial probability distribution</td>
<td></td>
</tr>
<tr>
<td>Stoichiometric and constrained-based modeling</td>
<td>does not need detailed kinetics</td>
<td>can lead to a large solution space</td>
<td>[147, 80]</td>
</tr>
<tr>
<td></td>
<td>easy to reconstruct for large networks</td>
<td>assumes optimality of metabolism</td>
<td></td>
</tr>
<tr>
<td>Nonlinear ordinary differential equations</td>
<td>detailed description of a metabolic network</td>
<td>large number of parameters</td>
<td>[80]</td>
</tr>
<tr>
<td>(kinetic models)</td>
<td></td>
<td>parameter estimation computationally costly</td>
<td></td>
</tr>
<tr>
<td>Stochastic modeling</td>
<td>accounts for cell-to-cell variability</td>
<td>computationally intensive</td>
<td>[168, 177]</td>
</tr>
<tr>
<td>Boolean networks</td>
<td>can be applied on qualitative experimental data</td>
<td>describes time as a discrete variable only qualitative information</td>
<td>[80]</td>
</tr>
<tr>
<td></td>
<td>computational tractability</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Qualitative differential equations</td>
<td>no stoichiometric or rate constants needed</td>
<td>only qualitative information</td>
<td>[80, 196]</td>
</tr>
<tr>
<td></td>
<td></td>
<td>difficult to apply to large networks</td>
<td></td>
</tr>
</tbody>
</table>

6.2.5 Software platforms and standards

For integration of different types of data or data from different laboratories, there is a demand for standards and storage formats. There is a need for a data and model storage format that can be imported in all software tools. Many current software tools support XML (eXtensible Markup Language) [196]. For different applications, different XML-compatible formats have to be developed. Examples are SBML (Systems Biology
Conclusion and outlook

Markup Language) for biochemical networks and models and CellML (Cellular Markup Language) for mathematical models [93]. Exchange of data and models is hampered by several factors. First, identical objects (e.g. genes, proteins, metabolites) and relationships (e.g. reactions) are named with synonyms. There is a need for standardized nomenclature [3]. Second, each discipline uses its own standards for data exchange. Examples are MIAME for microarrays, MIAPE for proteomics and MIAMET for metabolomics [155, 3]. Standards for interdisciplinary research have to be developed [155]. The MIBBI project (Minimal Information for Biological and Biomedical Investigation) attempts to reach this goals by incorporating links to standards in a portal website, searching for overlaps among different standards and determining areas where a uniform standard is most needed (e.g. description of the study design) [198].

6.2.6 Experimental design

The type of experiments needed is dependent on the goal of the study. Therefore, it is important that biologists and statisticians collaborate to determine the best experimental design for answering a specific research question [53]. For answering questions about the functioning of cellular systems, one has to measure and analyze internal metabolites, while for studying a whole organism also external metabolites are required [69]. For classification purposes, semi-quantitative data (metabolite fingerprinting) data are sufficient [53]. For supervised analysis, like correlation analysis and mutual information, it is important that a large number of replicates is measured [21]. For modeling dynamics and parameter estimation, time-resolved quantitative data are required [13]. For studying a few metabolites that are affected by perturbations, one needs to perform targeted analysis. For exploring pathways, metabolite profiling data are necessary [69, 53].
Acknowledgments

This project was financed by the Netherlands Metabolomics Centre (NMC), which is part of the Netherlands Genomics Initiative - Netherlands Organization for Scientific Research.
Summary

Summary for scientists

Metabolism is the whole of all chemical processes in an organism that enable adaptation to changing environments. The intermediates of metabolism, called metabolites, are organized in pathways, which are part of a large network. Analyzing how those pathways function is an important topic in systems biology, because it contributes to understanding cellular mechanisms. Metabolic networks are important for different disciplines. In medicine, they can help to distinguish healthy and diseased or study the effect of drugs. Metabolic networks can be used in plant biology and microbiology to study the effect of environmental perturbations, like availability of nutrient and carbon sources.

This thesis focuses on deriving metabolic network information from time-resolved metabolomics data. Pathways can be studied on different levels: one can derive the structure, reaction coefficients, directionality of the reactions or kinetic parameters. One can also compare networks between conditions to infer information about their regulation. Another type of studies focuses on adaptation of organisms to their environment. Pathways are no static entities, but are highly dynamic. This thesis focuses on inferring those dynamic properties from time-resolved metabolomics data.

Chapter 2 presents a study on deriving the structure and directionality of metabolic networks from time-resolved metabolomics data. Network inference methods are evaluated by using appropriate simulated data.
Summary

Current measurement methods are contrasted with computational methods. The results show large discrepancies between the requirements of computational methods and contemporary measurement practice.

In **chapter 3** Goeman’s global test, a statistical method from gene-expression analysis, is used to test metabolic pathway differences between conditions. Testing on experimental data of *E. coli* and *S. cerevisiae* shows that Goeman’s global test can be generalized to metabolomics.

In **chapter 4**, we show how correlation analysis in time-resolved metabolomics data under different conditions can give more insight in the regulation of biochemical processes. Correlation analysis is combined with *a priori* information about the reaction scheme to infer possible scenarios for the regulation of a pathway. These regulation scenarios are related to changes in the distribution of reaction rates.

**Chapter 5** focuses on integrating experimental data into Dynamic Flux Balance Analysis (DFBA), a method to study reaction rates over time. Combining DFBA with experimental data reduces both the solution space and the computational complexity of standard DFBA. Applications of the DFBA method are testing hypotheses about cellular adaptation to the environment and estimating reaction rate profiles. A case study about hypothesis testing is presented. Hypotheses about adaptation of *S. cerevisiae* to a glucose pulse are incorporated in the DFBA framework as an objective function and the resulting reaction rate profiles are confronted with the literature.

The thesis concludes with an overview of challenges for future research in metabolic network inference (**chapter 6**).
Summary

Summary for non-scientists

Metabolism is the whole of all chemical processes in a living organism. These processes ensure that the organism grows and is resistant to changes in the environment. Metabolites are grouped in sequences of subsequent metabolic processes, called metabolic pathways, which are connected to each other in a large network.

The study of metabolic pathways is important for various disciplines in scientific research and industry, including medicine, pharmacy and food industry.

In this thesis, metabolic pathways are studied by using data of metabolite concentrations which are measured at different time points. Time plays an important role in the study of metabolism because metabolic processes are very dynamic.

Chapter 2 examines mathematical methods for discovering new metabolic pathways. From the comparison of the requirements of the mathematical methods with current laboratory practice, we can conclude that mathematical and experimental methods are not consistent with each other.

In chapters 3 and 4 we compare metabolic pathways under different conditions. In chapter 5 we search for biological principles that ensure that certain metabolic processes change due to adaptation of the organism to the environment, while other necessary processes are maintained.

The thesis concludes with recommendations for future research (chapter 6).
Samenvatting

Samenvatting voor wetenschappers

Metabolisme is het geheel van alle chemische processen dat ervoor zorgt dat een organisme zich kan aanpassen aan een veranderende omgeving. De stoffen die deel uitmaken van het metabolisme, metabolieten, zijn gegroepeerd in metabole paden, die deel uitmaken van een groot netwerk. Analyseren hoe deze metabole paden functioneren is een belangrijk onderwerp in de systeembiologie, omdat het bijdraagt tot het begrijpen van cellulaire mechanismen. Metabole netwerken zijn belangrijk voor verschillende disciplines. In de geneeskunde kunnen ze bijdragen bij het onderscheid maken tussen gezond en ziek en bij het onderzoeken van het effect van geneesmiddelen. In plantbiologie en microbiologie kunnen metabole netwerken gebruikt worden om het effect van verstoringen in de omgeving te bestuderen, zoals de aanwezigheid van koolstofbronnen en nutriënten.

Dit proefschrift richt zich op het afleiden van informatie over metabole netwerken uit tijdsopgeloste data. Metabole paden kunnen bestudeerd worden op verschillende niveaus. Men kan de structuur van het netwerk afleiden, de reactiecoëfficiënten, de Richting van de reacties of kinetische parameters. Daarnaast kan men ook netwerken vergelijken onder verschillende condities om informatie af te leiden over regulatie. Een ander soort studies richt zich op het afleiden van principes die leiden tot adaptatie van organismen aan hun omgeving.

Metabole paden zijn geen statische entiteiten, maar zijn zeer dynamisch.
Samenvatting

Dit proefschrift richt zich op het afleiden van deze dynamische eigenschappen uit tijdsopgeloste metabole data.

**Hoofdstuk 2** beschrijft een studie over het schatten van de structuur van metabole netwerken en de richting van de reacties uit tijdsopgeloste metabole data. Methoden voor het schatten van netwerken worden geëvalueerd met behulp van geschikte gesimuleerde data. Huidige meet- en computationale methoden worden tegenover elkaar gesteld. De resultaten tonen tegenstrijdigheden tussen de vereisten van computationale methoden en huidige meetmethoden.

In **hoofdstuk 3** wordt Goeman’s global test, een statistische methode uit de gen-expressie analyse, gebruikt om te testen of er verschillen zijn in metabole paden tussen twee of meer condities. Het testen van deze methode op experimentele data van *E.coli* en *S.cerevisiae* toont aan dat Goeman’s global test veralgemeend kan worden naar de studie van het metaboloom.

In **hoofdstuk 4** tonen we hoe correlatieanalyse in tijdsopgeloste data onder verschillende condities meer inzicht kan geven in de regulatie van biologische processen. Correlatie-analyse wordt gecombineerd met voorkennis over het reactieschema om mogelijke regulatiescenario's voor een pathway af te leiden. Deze regulatiescenario's zijn gerelateerd met veranderingen in de distributie van reactiesnelheden.

**Hoofdstuk 5** richt zich op het integreren van experimentele data in Dynamische Flux Balans Analyse (DFBA), een methode om reactiesnelheden over de tijd te bestuderen. Het combineren van DFBA met experimentele data reduceert zowel de oplossingenverzameling als de computationale complexiteit van standaard DFBA. Toepassingen van de DFBA methode zijn het testen van hypothesen over hoe de cel zich aanpast aan haar omgeving en het schatten van reactiesnelheidsprofielen. We presenteren hier een gevalsstudie over het testen van hypothesen. Hypothesen over de aanpassing van *S.cerevisiae* bij een glucose puls worden ingebracht in de DFBA methode als doelstellingsfunctie en de resulterende reactiesnelheidsprofielen worden geconfronteerd met de literatuur.

Dit proefschrift besluit met een overzicht van uitdagingen voor toekomstig onderzoek over metabole netwerken (**hoofdstuk 6**).
Samenvatting

Samenvatting voor niet-wetenschappers

De stofwisseling, ook metabolisme genoemd, is het geheel van alle chemische processen in een levend organisme. Ze zorgen ervoor dat het organisme groeit en bestand is tegen veranderingen in de omgeving. Metabolieten zijn de stoffen die deel uitmaken van het metabolisme. Metabolieten zijn gegroepeerd in reeksen van opeenvolgende stofwisselingsprocessen, metabole paden, die met elkaar verbonden zijn tot een groot netwerk. Het bestuderen van metabole paden is belangrijk voor verschillende disciplines in het wetenschappelijk onderzoek en de industrie, onder meer in de geneeskunde, de farmacie en de voedingsindustrie. In dit proefschrift worden metabole paden bestudeerd met behulp van data van metaboliet concentraties die gemeten werden op verschillende tijdstippen. Tijd speelt een belangrijke rol in het onderzoek van het metabolisme omdat stofwisselingsprocessen zeer dynamisch zijn.

Hoofdstuk 2 bestudeert wiskundige methoden voor het ontdekken van nieuwe metabole paden. Uit vergelijking van de vereisten voor het toepassen van deze wiskundige methoden met de mogelijkheden van de huidige meettechnieken kunnen we besluiten dat deze niet op elkaar zijn afgestemd.

In hoofdstuk 3 en 4 vergelijken we metabole paden onder verschillende condities. In hoofdstuk 5 gaan we op zoek naar de biologische principes die ervoor zorgen dat bepaalde stofwisselingsprocessen zich aanpassen aan de omgeving, terwijl andere noodzakelijke processen in stand gehouden worden.

Het proefschrift wordt afgesloten met aanbevelingen voor verder onderzoek (hoofdstuk 6).
Acknowledgments

This thesis would not have been possible without the support and help of many people.
I wish to thank, first and foremost, my promotor Age Smilde for his inspiring discussions about network inference, carefully reading of my manuscripts and giving suggestions for improving my research.
It gives me great pleasure in acknowledging the excellent supervision of my PhD research and the support of my co-promotores Huub Hoefsloot and Margriet Hendriks.
I am indebted to my many colleagues, former colleagues, guests and students in the BDA group who supported me during my PhD project: Age, Antoine, Huub, Johan, Gooitzen, Andrew, Chengjian, Daniël, Edoardo, Ewa, Jeroen, Marcel, Oxana, Suzanne, Maikel, Dicle, Ewoud, Joe, Kilian, Mateusz, Polina, Ishtiaq, Eelke, Jack, Serge, Siemen, Tim, Velitchka, Zha Ying, Iven, José Maria, Mari, Samuel, Xiang, Bastiaan, Eva and Maarten.
Jildau Bouwman (TNO) is gratefully acknowledged for bringing me into contact with André Canelas, with whom I had nice collaborations during my PhD project.
It is with immense gratitude that I acknowledge the following people for providing me with data sets: André Canelas (TU Delft / DSM), Mariët van der Werf (TNO / DSM) and Peter Punt (TNO).
I would like to thank Paul Eilers (Erasmus MC) for his support about methods for smoothing. I gratefully acknowledge Bas Teusink, Frank Bruggeman and Timo Maarleveld (VU) for their input in the DFBA
study. I also want to thank Bas Teusink for his input in the correlation paper. I am grateful to Gertien Smits (UvA) for giving suggestions for objective functions in the DFBA study. I owe my deepest gratitude to Karen van Eunen and Barbara Bakker (UMCG) for providing me with the glucose pulse model.

Daniël Vis and Johan Andriessen are gratefully acknowledged for being my paranymphs.

I cannot find words to express my gratitude to my parents, my partners family and my friends for the mental support during my PhD project. Special thanks also to Ad and Ans for providing me with accommodation during the last 3.5 months of my PhD.

Finally, I would like to thank my partner Johan for all his patience during my PhD project and his mental support during difficult periods.
Publications


Bibliography


Bibliography


164


Bibliography


[168] H. M. Sauro, A. M. Uhrmacher, D. Harel, M. Hucka, M. Kwiatkowska, P. Mendes, C. A. Shaffer, L. Stromback, and


Bibliography


