Network inference from time-resolved metabolomics data

Diana M. Hendrickx

Network inference from time-resolved metabolomics data

ACADEMISCH PROEFSCHRIFT

ter 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

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.

Table 1.1: Overview of metabolic control mechanisms.

mechanism	**definition**
enzyme induction	increased enzyme synthesis in the presence or absence of a certain metabolite [73, 126]
enzyme repression	decreased enzyme synthesis in the presence or absence of a certain metabolite [73, 126]
substrate-level control	high levels of product inhibit the substrate to react [126]
feedback control	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)
feed forward control	a metabolite activates or inhibits a further step in the pathway [196, 126] (see Figure 1.1 c and d)

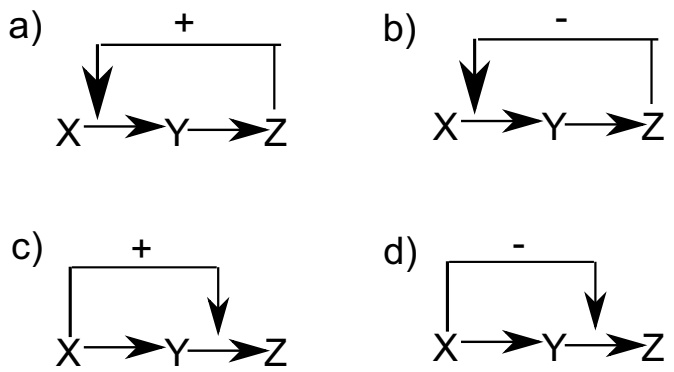

Figure 1.1: Metabolic control mechanisms a) positive feedback control b) negative feedback control c) positive feedforward control d) negative feed forward control

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].

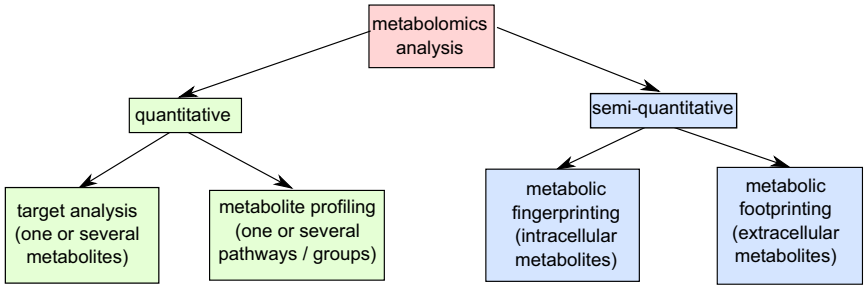

Figure 1.2: Types of measurements in metabolomics.

1.1.2 Sampling methods for time-resolved microbial metabolomics

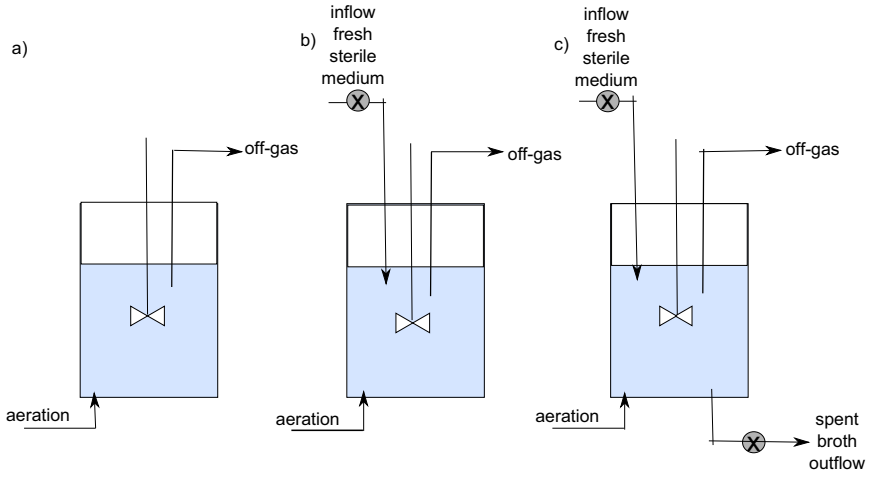

Figure 1.3: Operation modes of a bioreactor: a) batch b) fed-batch c) continuous. Adapted from Mashego *et al*[124].

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)).

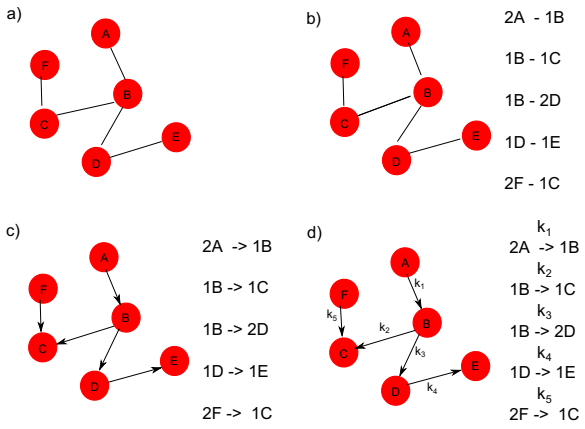

Figure 1.4: Different levels of studying metabolic pathways. a) topology b) stoichiometry c) directionality d) kinetics

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.

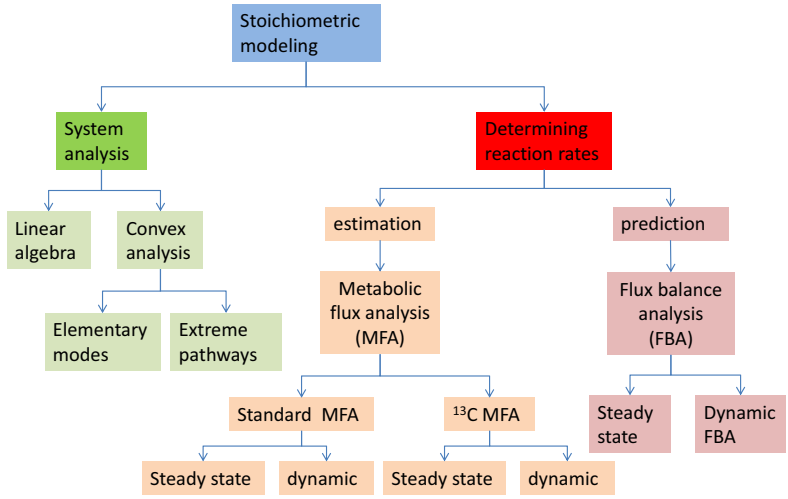

Figure 1.5: Current methodologies in stoichiometric modeling.

Stoichiometric models make use of mass balances $dC/dt = S \cdot v$, 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

[232]. 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[1]

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

[1]This chapter is based on Diana M. Hendrickx, Margriet M. W. B. Hendriks, Paul H. C. Eilers, Age K. Smilde and Huub C. J. Hoefsloot (2011). Reverse engineering of metabolic networks, a critical assessment. Mol. BioSyst, Volume 7:2 (2011) pages 511-520

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 sub second 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 Çakìr and coworkers [27] that uses partial Pearson correlations (also called a Graphical Gaussian Model) is discussed. Secondly, a correlation

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.

Sampling method	**Organism**	**Sampling frequency**	**Number of components**	**Analytical methods and measurement error**	**References**
fast sampling from bioreactor (stopped-flow technique)	micro-organisms (*E.coli*, yeast)	sub second	5-15 metabolites	enzymatic analysis; HPLC; variation between duplicates $< 3\%$; variation between different extractions $< 10\%$;	[40, 20] [210]

fast sampling from bioreactor (automated sampling device)	micro-organisms (*E.coli*)	sub second	15-30 metabolites	enzymatic analysis; HPLC; LC-MS-MS; RSD $= 5 - 10\%$	[169] [18] [210]
fast sampling from bioreactor (Bioscope)	micro-organisms (yeast)	seconds	5-20 metabolites	enzymatic analysis; LC-MS-MS; UV-VIS; GC-FID; LC-ESI-MS-MS (IDMS); RSD $= 5 - 25\%$	[124] [217] [125] [231] [104] [106] [210]
fast sampling from bioreactor (stimulus response approach)	micro-organisms (yeast)	seconds	5-20 metabolites	enzymatic analysis; HPLC; bio-luminescence; variation between triplicates $< 5\%$; variation between different extractions $< 10\%$; error bio-luminescence $\geq 20\%$;	[203] [202] [210]

				error HPLC $\leq 6\%$	
slow sampling from bioreactor	micro-organisms (yeast)	hours	10-15 metabolites	LC-ESI-MS-MS (IDMS); $= 10 - 30\%$	[231]
slow sampling from bioreactor	micro-organisms (*E.coli, B.subtilis, P.freuden-reichii*)	hours	tens or hundreds metabolites	GC-MS; IP-LC-MS; OS-GC-MS; RSD $= 5 - 20\%$	[94] [164]
tissue samples	plants	minutes/ hours	15-20 metabolites	GC-MS; LC-MS; NMR; RSD $= 10 - 20\%$	[12] [158]
blood samples	human (males)	10 minutes; leptin: 20 minutes	1-8 hormones	immunoassay's; inter assay variation $< 20\%$; intra-assay variation $< 5\%$; RSD $= 5 - 25\%$	[148] [86, 98] [101] [99] [100]
blood samples	dog, rat (liver)	45 minutes	4 lipids	GC-MS; RSD $= 5 - 10\%$	[8]
blood samples	human (females)	minutes/ hours	5-10 hormones	immunoassay's; inter assay variation $< 15\%$; intra-assay variation	[97, 96]

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

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.

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

[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]:

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

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

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

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

Çakìr 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]. Çakìr 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

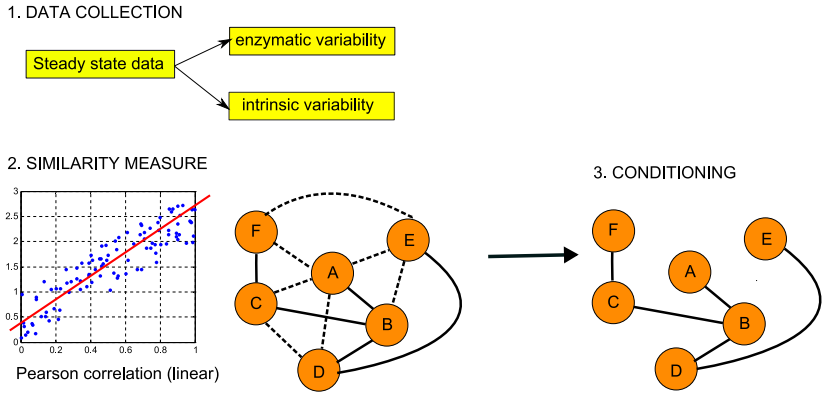

Figure 2.1: Illustration of the reverse engineering method based on partial Pearson correlation. Adapted from Çakìr and coworkers [27].

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]:

$$\frac{dS}{dt} \approx J \cdot (S - S_0) \tag{2.3}$$

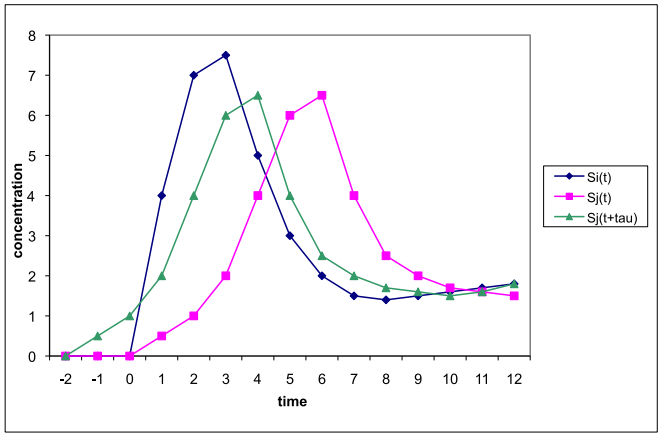

Figure 2.2: Illustration of time-lagged correlation. The time series $S_i\,(t)$ (blue) and $S_j\,(t)$ (magenta) show weak correlation. However, if $S_i\,(t)$ is compared with the time-lagged data $S_j\,(t+\tau)$ (green) for $\tau = 2$, a strong correlation is observed. Adapted from Crampin and coworkers [37].

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 δ needs to be specified. If an entry in the Jacobian is smaller than δ, 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

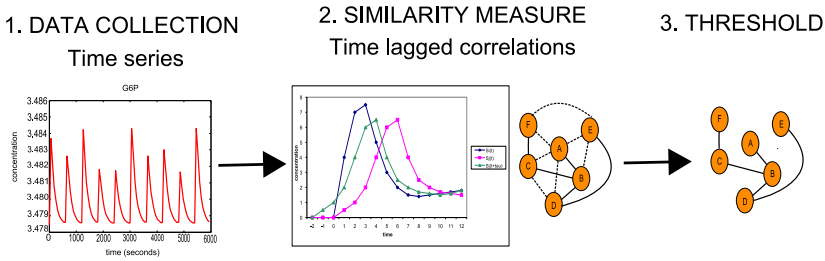

Figure 2.3: Overview of the reverse engineering method based on time-lagged correlations

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

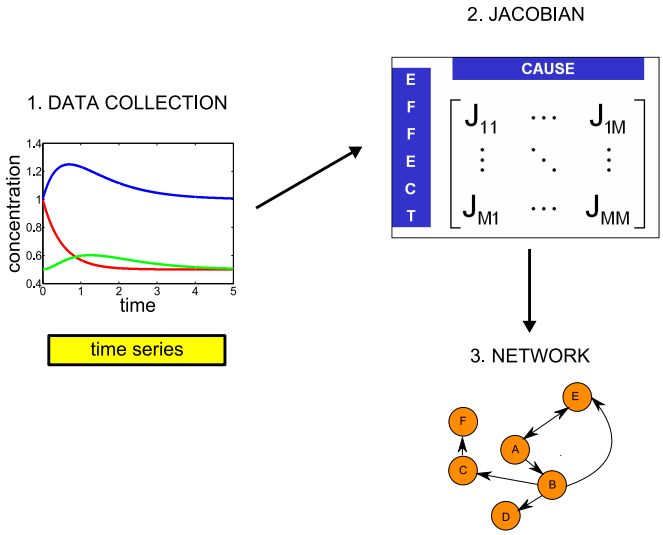

Figure 2.4: Overview of the Jacobian method.

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]:

$$\begin{aligned} g-score &= \sqrt{sensitivity \times specificity} \\ &= \sqrt{TPR \times TNR} \end{aligned}$$

The g-score is always a number between 0 and 1, where a g-score of 1 corresponds with perfect inference.

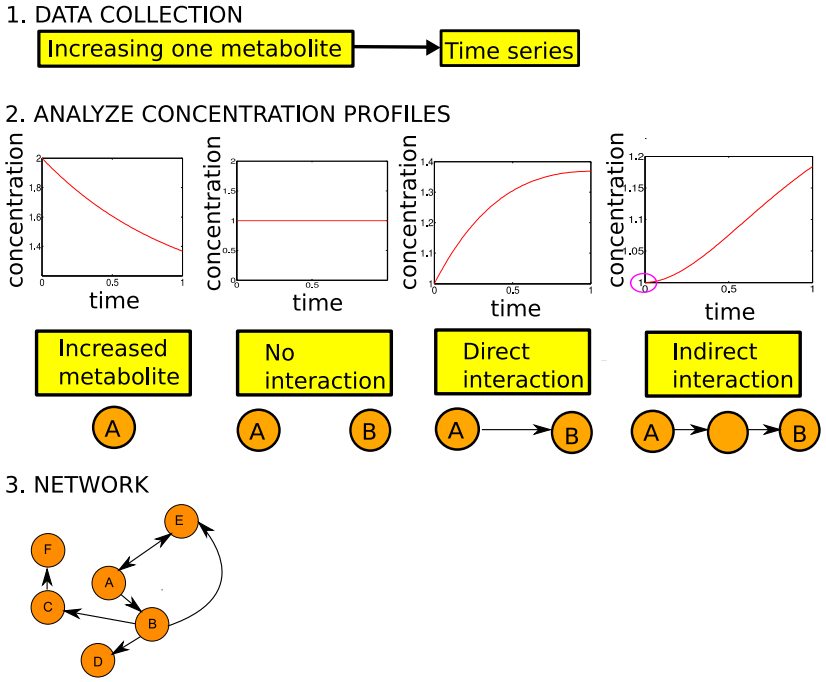

Figure 2.5: Overview of the zero slopes method.

The information in the appendix of Çakìr 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 Chassagnole and coworkers [30] for *E. coli*.
Simulations for the method based on similarity measures are performed like described in Çakìr 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 equilibrum 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

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 2.3: Results of the four methods for *S. cerevisiae* and *E. coli*

		S. cerevisiae			***E. coli***	
method[b]	**samples/Δt**[c]	**noiseless g-score (TPR,TNR)**[d]	**3% noise g-score**	**samples/Δt**	**noiseless g-score (TPR,TNR)**	**3% noise g-score**
partial Pearson correlations	enzymatic variability[e]	0.75 (0.69,0.81)[f]	0.63	enzymatic variability	0.74 (0.66,0.84)[g]	0.49
	intrinsic variability[h]	0.82 (0.77,0.87)[i]	0.52	intrinsic variability	0.75 (0.61,0.92)[j]	0.55
time-lagged correlations	Δt = 0.2 min[l]	0.63 (0.48,0.82)	0.32	Δt = 120 s[k]	0.61 (0.46,0.81)	0.42
	Δt = 0.1 min	0.63 (0.48,0.84)	0.46	Δt = 60 s	0.65 (0.51,0.83)	0.34

[b] *For details about the four approaches see Methods section.*

[c] *Δt is the time between two measurements.*

[d] *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 Çakìr and coworkers [27].*

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

[h] *See Methods.*

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

[j] *Results for noiseless data based on the appendix of Çakìr 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.*

penalized	Δt	0.33	0.27	Δt	0.55	0.04
Jacobian	$= 10^{-3}$ min	(0.46,0.71)		$= 10^{-2}$ s	(0.73,0.75)	
method	Δt	0.49	0.15	Δt	0.69	0.04
	$= 10^{-4}$ min	(0.51,0.95)		$= 10^{-3}$ s	(0.92,0.75)	
zero slopes	Δt	0.65	0.48	Δt	0.88	0.61
method	$= 10^{-3}$ min	(0.95,0.44)		$= 10^{-2}$ s	(0.83,0.93)	
	Δt	0.77	0.57	Δt	0.91	0.62
	$= 10^{-4}$ min	(1,0.59)		$= 10^{-3}$ s	(0.85,0.97)	

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

Property	**Partial Pearson correlations**	**Time-lagged correlations**	**Penalized Jacobian method**	**Zero slopes method**
Distinguish between direct and indirect interactions	reasonable	reasonable	for large sampling frequencies	for large sampling frequencies
Influence of noise	large	large	very large	large

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 Çakìr [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 10^3, 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 occured. 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]:

$$\rho_{ij} = \sum_{k=1}^{m} \left(\frac{S_i(k) - \bar{S}_i}{\sigma_i} \right) \left(\frac{S_j(k) - \bar{S}_j}{\sigma_j} \right) \tag{2.4}$$

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; σ_i and σ_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 Π can be calculated from the matrix P with the following two formulas [170]:

$$\Omega = P^{-1} = (\omega_{ij}) \tag{2.5}$$

$$\Pi_{ij} = \frac{-\omega_{ij}}{\sqrt{\omega_{ii} \cdot \omega_{jj}}} \tag{2.6}$$

A significance measure for the similarity scores was calculated by performing a permutation test like described by Çakìr 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]:

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

if $\tau \geq 0$

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

if $\tau < 0$
where $i, j \in \{1, \cdots, n\}$ 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 ϕ_{ii} and ϕ_{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]:

$$c_{ij} = max_{\frac{1-m}{2} \leq \tau \leq \frac{m-1}{2}} |C_{ij}(\tau)| \tag{2.7}$$

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

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]:

$$\frac{dS}{dt} \approx J \cdot (S - S_0) \tag{2.8}$$

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

$$\frac{dS(t_n)}{dt} \approx \frac{-S(t_n + 2\Delta t)}{12 \cdot \Delta t} + \frac{8S(t_n + \Delta t)}{12 \cdot \Delta t} - \frac{8S(t_n - \Delta t)}{12 \cdot \Delta t} + \frac{S(t_n - 2\Delta t)}{12 \cdot \Delta t}$$

where $S(t_n)$ is the concentration at time $t = t_n$ and Δt 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

$$\left\| XJ^T - Y \right\|_2^2$$

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:

$$J^T = \left(X^T X\right)^{-1} X^T Y$$

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 γ is first added to the diagonal of $X^T X$, before the matrix is inverted:

$$J^T = \left(X^T X + \gamma I\right)^{-1} X^T Y$$

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]:

$$\left\| XJ^T - Y \right\|_2^2 + \lambda \left\| J^T \right\|_1 \tag{2.9}$$

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

$$j_{new} = \left(X^T X + \Gamma^T \Gamma\right)^{-1} X^T y \tag{2.10}$$

where $y = vec(Y)$ and

$$\Gamma^T \Gamma = \lambda^2 \begin{bmatrix} \frac{1}{\epsilon + |j_1| + \kappa j_1^2} & 0 & \cdots & 0 \\ 0 & \ddots & 0 & \vdots \\ \vdots & 0 & \ddots & 0 \\ 0 & \cdots & 0 & \frac{1}{\epsilon + |j_n| + \kappa j_n^2} \end{bmatrix} \tag{2.11}$$

where ϵ is a very small number $\left(10^{-6} - 10^{-4}\right)$ and λ and κ 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 $k_2 = k_4 = k_6 = 2$.

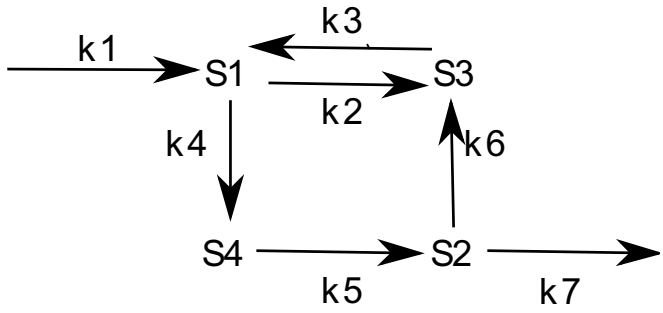

Figure 2.6: A hypothetical network, consisting of four metabolites.

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.

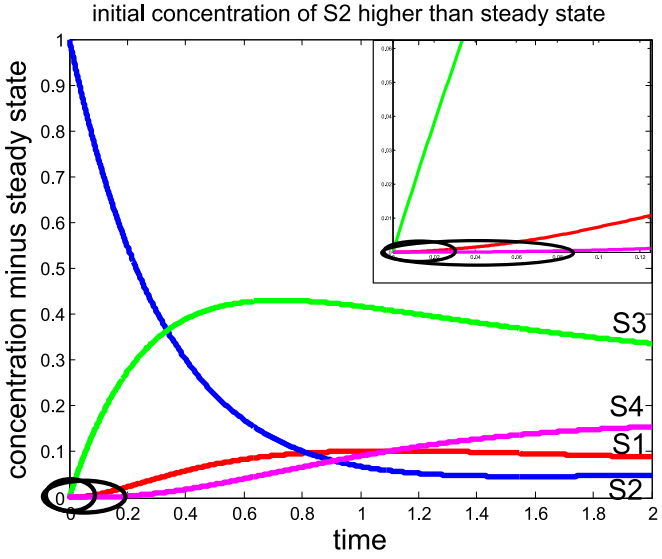

Figure 2.7: Concentration profiles for the system in Figure 2.6 when the initial concentration of S_2 is higher than steady state and all the other initial concentrations are at steady state. Zero slopes are indicated with black circles.

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 η has to be specified. Initial slopes smaller than η 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 Çakìr is gratefully acknowledged for providing us with Matlab code on partial Pearson correlations and power point 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.

Chapter 3

Global test for metabolic pathway differences between conditions[2]

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-

[2]This chapter is based on Diana M. Hendrickx, Huub C.J. Hoefsloot, Margriet M.W.B. Hendriks, André B. Canelas and Age K. Smilde (2012). Global test for metabolic pathway differences between conditions. Analytica Chimica Acta, Volume 719, pages 8-15

ronmental conditions. 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

[179].
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

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)).

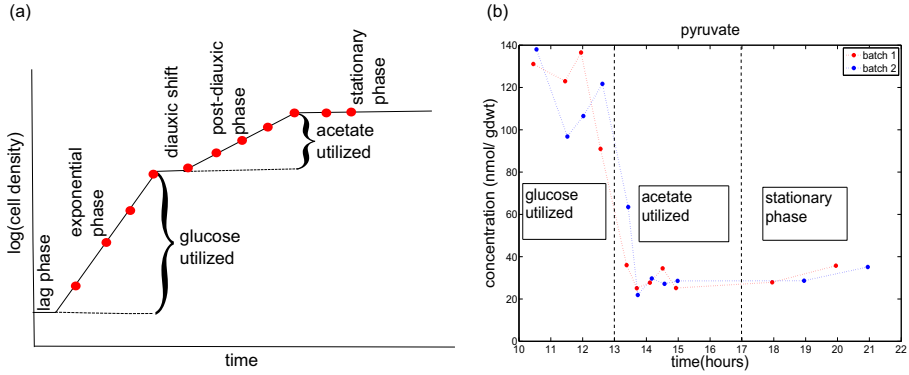

Figure 3.1: (a) Diauxic growth curve. The red points indicate in which phases the measurements were taken. (b) An example of a metabolite concentration profile (pyruvate) under diauxic growth. The different growth phases are indicated on the graph. Abbreviations: nmol, nanomoles; gdwt, gram dry weight.

3.2.2 *Saccharomyces cerevisiae* data set

LC-MS data [25, 140, 81] of continuous cultures [1] 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.

[1] chemostat cultures, continuous inflow and outflow

Table 3.1: Description of the experiments.

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

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]:

$$E\left(Y_i|\beta\right) = h^{-1}\left(\alpha + \sum_{j=1}^{m} x_{ij}\beta_j\right) \tag{3.1}$$

where α is the intercept, β_j the regression coefficient for metabolite j and h the logit function [130]:

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

where $\mu_i = E\left(Y_i|\beta\right)$ $(i = \{1, \cdots, n\})$. The regression coefficients β_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 β_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.

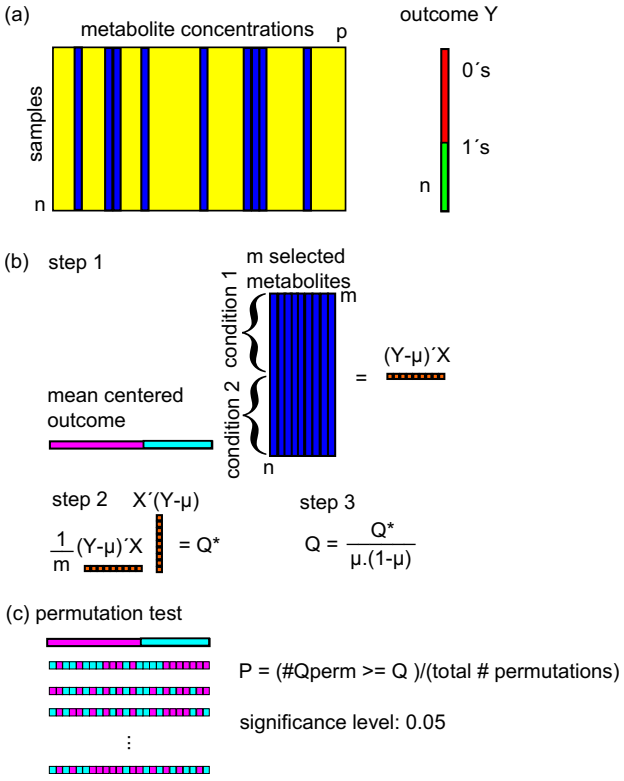

Figure 3.2: Overview of the score test. (a) From the autoscaled data matrix, m metabolites belonging to the same pathway are selected. A binary outcome is defined. (b) A score statistic Q is calculated from the mean centered outcome and the matrix of selected metabolites. (c) The significance of the relation between the group of metabolites (pathway) and the outcome is determined by performing a permutation test.

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 β_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 β_j's are drawn from some common distribution. This distribution is characterized by $E(\beta) = 0$ and $cov(\beta) = E(\beta\beta') = \tau^2 I_m$ where τ is a constant number and I_m the identity matrix of size m [60, 174]. The value τ^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 $cov(r) = E(rr') = E(X\beta\beta'X') = XE(\beta\beta')X' = X\tau^2X' = \tau^2XX'$ [59, 174]. Thus r is a random variable with expectation zero and a covariance containing the parameter τ^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]:

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

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 α is unknown, the exact values of μ and μ_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 $\hat{\mu_{2i}} =$

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

$$Q = \frac{(Y-\hat{\mu})'\,R\,(Y-\hat{\mu})}{\hat{\mu_2}} \tag{3.3}$$

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), step1, $(Y-\mu)'\,X$ 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.

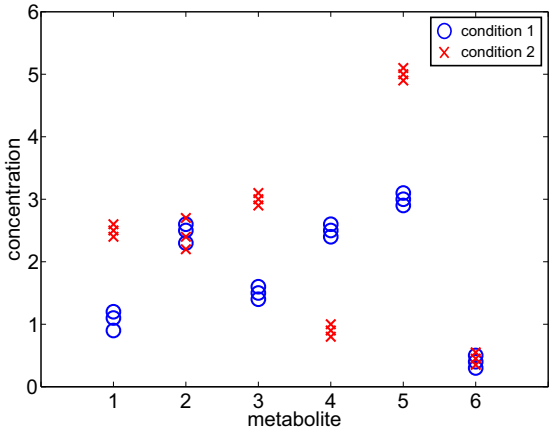

Figure 3.3: An illustration of the type of differences detected by Goeman's global test. In this (hypothetical) example, the concentration levels of a group of six metabolites are shown for six samples. Three samples are measured under condition 1 (blue circles) and another three samples under condition 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.

KEGG pathway	**number of metabolites in the group**
glycolysis + branches	7
TCA cycle	8
alanine, aspartate and glutamate metabolism	5
glycine, serine and threonine metabolism	4
valine, leucine and isoleucine degradation	4
valine, leucine and isoleucine biosynthesis	5
purine metabolism	5

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.

Pathway	**Q statistic**	**not adjusted p-value**	**Bonferroni adjusted p-value**
glycolysis + branches	200	$< 10^{-4}$	**0.0002**
TCA cycle	135	0.0001	**0.0008**
alanine, aspartate and glutamate metabolism	152	$< 10^{-4}$	**0.0002**
glycine, serine and threonine metabolism	126	$< 10^{-4}$	**0.0002**
valine, leucine and isoleucine degradation	163	$< 10^{-4}$	**0.0002**
valine, leucine and isoleucine biosynthesis	163	$< 10^{-4}$	**0.0002**
purine metabolism	164	0.0004	**0.0027**

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.

outcome (perturbation 1/ perturbation 2)	**pathway**	**number of metabolites**	**total number of samples**	**number of samples pert.1**	**number of samples pert.2**
aerobic / anaerobic	glycolysis	6	15	11	4
	TCA cycle	7	15	11	4
glucose pulse (aerobic) / glucose deprivation (aerobic)	glycolysis	6	11	9	2
	TCA cycle	7	11	9	2
glucose pulse of 10 mM (aerobic) / glucose pulse $\leq$ 3 mM (aerobic)	glycolysis	6	11	4	7
	TCA cycle	7	11	4	7

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

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.

pathway	**time point**	**Q statistic**	**not adjusted p-value**	**Bonferroni adjusted p-value**
glycolysis	steady state	114	0.001	**0.002**
	10 s	81	0.001	**0.002**
	1 min	79	0.001	**0.002**
	2 min	83	0.002	**0.003**
TCA cycle	steady state	166	0.001	**0.002**
	10 s	147	0.001	**0.002**
	1 min	128	0.001	**0.002**
	2 min	106	0.001	**0.002**

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

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.

outcome (perturbation 1 / perturbation 2)	**pathway**	**time point**	**Q statistic**	**not adjusted p-value**	**Bonferroni adjusted p-value**
glucose pulse/glucose deprivation	glycolysis	10 s	18	0.036	0.073
		1 min	53	0.018	**0.036**
		2 min	51	0.018	**0.036**
	TCA cycle	10 s	3	0.273	0.545
		1 min	18	0.018	**0.036**
		2 min	27	0.018	**0.036**
glucose pulse of 10 mM/glucose pulse $\leq$ 3 mM	glycolysis	10 s	31	0.003	**0.006**
		1 min	11	0.118	0.236
		2 min	10	0.112	0.224
	TCA cycle	10 s	4	0.115	0.230
		1 min	8	0.112	0.224
		2 min	5	0.430	0.861

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 can not 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 ($\leq$ 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.

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.

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.

data set	**Q statistic**	**p-value**
FBP + 5 columns of random numbers	41.0395	**0.003**
FBP + 20 columns of random numbers	21.3772	**0.046**
FBP + 30 columns of random numbers	19.9447	0.081

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.

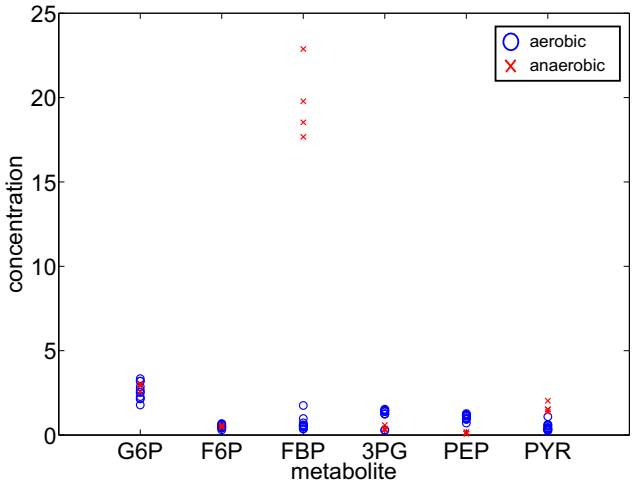

Figure 3.4: Steady state concentration levels of metabolites of glycolysis in the *S. cerevisiae* data set under aerobic (blue circles) and anaerobic (red x-marks) glucose-limited conditions. Abbreviations: G6P, glucose-6-phosphate; F6P, fructose-6-phosphate; FBP, fructose-1,6-bisphosphate; 3PG, 3-phosphoglycerate;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

metabolomics.

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 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].

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.

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

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

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.

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 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.

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 computationele methoden worden tegenover elkaar gesteld. De resultaten tonen tegenstrijdigheden tussen de vereisten van computationele 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 regulatiescenarios voor een pathway af te leiden. Deze regulatiescenarios 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 computationele 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 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

Diana M. Hendrickx, Margriet M. W. B. Hendriks, Paul H. C. Eilers, Age K. Smilde and Huub C. J. Hoefsloot (2011). Reverse engineering of metabolic networks, a critical assessment. Mol. BioSyst, Volume 7:2 (2011) pages 511-520

Diana M. Hendrickx, Huub C.J. Hoefsloot, Margriet M.W.B. Hendriks, André B. Canelas and Age K. Smilde (2012). Global test for metabolic pathway differences between conditions. Analytica Chimica Acta, Volume 719, pages 8-15

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

Bibliography

[1] M. S. Abu-Asab, M. Chaouchi, S. Alesci, S. Galli, M. Laassri, A. K. Cheema, F. Atouf, J. VanMeter, and H. Amri. Biomarkers in the age of omics: time for a systems biology approach. *OMICS*, 15(3):105–112, 2011.

[2] A. Agrawal and A. Mittal. A dynamic time-lagged correlation based method to learn multi-time delay gene networks. *World Academy of Science, Engineering and Technology*, 9:167–174, 2005.

[3] C. H. Ahrens, U. Wagner, H. K. Rehrauer, C. Turker, and R. Schlapbach. Current challenges and approaches for the synergistic use of systems biology data in the scientific community. *EXS*, 97:277–307, 2007.

[4] B. Alberts, A. Johnson, J. Lewis, M. Raff, K. Roberts, and P. Walter. *Molecular Biology of the Cell. Fifth Edition.* Taylor & Francis, Inc., 2007.

[5] A. Arkin and J. Ross. Statistical construction of chemical-reaction mechanisms from measured time-series. *Journal of Physical Chemistry*, 99(3):970–979, 1995.

[6] A. Arkin, P. D. Shen, and J. Ross. A test case of correlation metric construction of a reaction pathway from measurements. *Science*, 277(5330):1275–1279, 1997.

[7] G. Balazsi, A. van Oudenaarden, and J. J. Collins. Cellular decision making and biological noise: from microbes to mammals. *Cell*, 144(6):910–925, Mar 2011.

[8] I. R. Bederman, A. E. Reszko, T. Kasumov, F. David, D. H. Wasserman, J. K. Kelleher, and H. Brunengraber. Zonation of labeling of lipogenic acetyl-coa across the liver - implications for studies of lipogenesis by mass isotopomer analysis. *Journal of Biological Chemistry*, 279(41):43207–43216, 2004.

[9] L. R. Berger and J. C. Hsu. Bioequivalence trials, intersection-union tests and equivalence confidence sets. *Statistical Science*, 11(4):283–302, 1996.

[10] R. L. Berger. Multiple hypothesis testing and acceptance sampling. *Technometrics*, 24(4):295–300, 1982.

[11] S. Bijlsma, L. Bobeldijk, E. R. Verheij, R. Ramaker, S. Kochhar, I. A. Macdonald, B. van Ommen, and A. K. Smilde. Large-scale human metabolomics studies: A strategy for data (pre-) processing and validation. *Analytical Chemistry*, 78(2):567–574, 2006.

[12] J. Boatright, F. Negre, X. L. Chen, C. M. Kish, B. Wood, G. Peel, I. Orlova, D. Gang, D. Rhodes, and N. Dudareva. Understanding in vivo benzenoid metabolism in petunia petal tissue. *Plant Physiology*, 135(4):1993–2011, 2004.

[13] R. Bonneau. Learning biological networks: from modules to dynamics. *Nat Chem Biol*, 4(11):658–664, Nov 2008.

[14] A. Boorsma, B. C. Foat, D. J. Vis, F. Klis, and H.J. Bussemaker. T-profiler: scoring the activity of predefined groups of genes using gene expression data. *Nucleic Acids Research*, 33:W592–W595, 2005.

[15] M. J. Brauer, J. Yuan, B. D. Bennet, W. Lu, E. Kimball, D. Botstein, and J. D. Rabinowitz. Conservation of the metabolomic response to starvation across two divergent microbes. *Proc Natl Acad Sci U S A*, 103(51):19302–19307, 2006.

[16] R. G. Brereton. *Applied chemometrics for scientists*. John Wiley & Sons Ltd, Chichester, West Sussex, England, 2007.

[17] F. J. Bruggeman and H. V. Westerhoff. The nature of systems biology. *Trends Microbiol*, 15(1):45–50, Jan 2007.

[18] A. Buchholz, J. Hurlebaus, C. Wandrey, and R. Takors. Metabolomics: quantification of intracellular metabolite dynamics. *Biomolecular Engineering*, 19(1):5–15, 2002.

[19] J. M. Buscher, D. Czernik, J. C. Ewald, U. Sauer, and N. Zamboni. Cross-platform comparison of methods for quantitative metabolomics of primary metabolism. *Analytical Chemistry*, 81(6):2135–2143, 2009.

[20] S. Buziol, I. Bashir, A. Baumeister, W. Claassen, N. Noisommit-Rizzi, W. Mailinger, and M. Reuss. New bioreactor-coupled rapid stopped-flow sampling technique for measurements of metabolite dynamics on a subsecond time scale. *Biotechnology and Bioengineering*, 80(6):632–636, 2002.

[21] D. Camacho, A. de la Fuente, and P. Mendes. The origin of correlations in metabolomics data. *Metabolomics*, 1(1):53–65, 2005.

[22] C. Camarasa, J. P. Grivet, and S. Dequin. Investigation by 13c-nmr and tricarboxylic acid (tca) deletion mutant analysis of pathways for succinate formation in saccharomyces cerevisiae during anaerobic fermentation. *Microbiology-Sgm*, 149:2669–2678, 2003.

[23] A. B. Canelas, C. Ras, A. ten Pierick, W.M. van Gulik, and J. J. Heijnen. An in-vivo data-driven framework for classification and quantification of enzyme kinetics and determination of apparent thermodynamic data. *Metabolic Engineering*, 13(3):294–306, 2011.

[24] A. B. Canelas, A. ten Pierick, C. Ras, R. M. Seifar, J. C. van Dam, W. M. van Gulik, and J. J. Heijnen. Quantitative evaluation of intracellular metabolite extraction techniques for yeast metabolomics. *Anal Chem*, 81(17):7379–7389, 2009.

[25] A. B. Canelas, W. M. van Gulik, and J. J. Heijnen. Determination of the cytosolic free nad/nadh ratio in saccharomyces cerevisiae under steady-state and highly dynamic conditions. *Biotechnology and Bioengineering*, 100(4):734–743, 2008.

[26] R. Caspi, T. Altman, K. Dreher, C. A. Fulcher, P. Subhraveti, I. M. Keseler, A. Kothari, M. Krummenacker, M. Latendresse, L. A. Mueller, Q. Ong, S. Paley, A. Pujar, A. G. Shearer, M. Travers, D. Weerasinghe, P. Zhang, and P. D. Karp. The metacyc database of metabolic pathways and enzymes and the biocyc collection of pathway/genome databases. *Nucleic Acids Res*, 40(Database issue):D742–D753, Jan 2012.

[27] T. Çakìr, M.M.W.B. Hendriks, J.A. Westerhuis, and A.K. Smilde. Metabolic network discovery through reverse engineering of metabolome data. *Metabolomics*, 5(3):318–329, 2009.

[28] T. Çakìr, K. R. Patil, Z. I. Onsan, K. O. Ulgen, B. Kirdar, and J. Nielsen. Integration of metabolome data with metabolic networks reveals reporter reactions. *Molecular Systems Biology*, 2:50, 2006.

[29] M. Chagoyen and F. Pazos. Mbrole: enrichment analysis of metabolome data. *Bioinformatics*, 27(5):730–731, 2011.

[30] C. Chassagnole, N. Noisommit-Rizzi, J. W. Schmid, K. Mauch, and M Reuss. Dynamic modeling of the central carbon metabolism of escherichia coli. *Biotechnol Bioeng*, 79(1):53–73, 2002.

[31] J. J. Chen, T. Lee, R. R. Delongchamp, T. Chen, and C. A. Tsai. Significance analysis of groups of genes in expression profiling studies. *Bioinformatics*, 23(16):2104–2112, 2007.

[32] X. Chen, A. P. Alonso, D. K. Allen, J. L. Reed, and Y. Shachar-Hill. Synergy between (13)c-metabolic flux analysis and flux balance analysis for understanding metabolic adaptation to anaerobiosis in e. coli. *Metab Eng*, 13(1):38–48, Jan 2011.

[33] X. Chen and Y. Shachar-Hill. Insights into metabolic efficiency from flux analysis. *J Exp Bot*, 63(6):2343–2351, Mar 2012.

[34] I. C. Chou and E. O. Voit. Estimation of dynamic flux profiles from metabolic time series data. *BMC Syst Biol*, 6(1):84, 2012.

[35] S. Cortassa and M. A. Aon. Metabolic control analysis of glycolysis and branching to ethanol-production in chemostat cultures of saccharomyces-cerevisiae under carbon, nitrogen, or phosphate limitations. *Enzyme and Microbial Technology*, 16(9):761–770, 1994.

[36] M. W. Covert and B. O. Palsson. Transcriptional regulation in constraints-based metabolic models of escherichia coli. *Journal of Biological Chemistry*, 277(31):28058–28064, 2002.

[37] E. J. Crampin, S. Schnell, and P. E. McSharry. Mathematical and computational techniques to deduce complex biochemical reaction mechanisms. *Progress in Biophysics and Molecular Biology*, 86(1):77–112, 2004.

[38] T. Dandekar, A. Fieselmann, S. Majeed, and Z. Ahmed. Software applications toward quantitative metabolic flux analysis and modeling. *Brief Bioinform*, page doi 10.1093/bib/bbs065, Nov 2012.

[39] R.A. De Graaf. *In vivo NMR spectroscopy: principles and techniques, 2nd edition.* John Wiley & Sons Ltd, Chichester, West Sussex, England, 2008.

[40] W. de Koning and K. van Dam. A method for the determination of changes of glycolytic metabolites in yeast on a subsecond time scale using extraction at neutral ph. *Anal Biochem*, 204(1):118–123, 1992.

[41] E. Delgado-Eckert. Reverse engineering time discrete finite dynamical systems: a feasible undertaking? *PloS one*, 4(3):e 4939, 2009.

[42] K. Van Deun, H. Hoijtink, L. Thorrez, L. Van Lommel, F. Schuit, and I. Van Mechelen. Testing the hypothesis of tissue selectivity:

the intersection-union test and a bayesian approach. *Bioinformatics*, 25(19):2588–2594, Oct 2009.

[43] D. Diez, A. M. Wheelock, S. Goto, J. Z. Haeggstrom, G. Paulsson-Berne, G. K. Hansson, U. Hedin, A. Gabrielsen, and C. E. Wheelock. The use of network analyses for elucidating mechanisms in cardiovascular disease. *Molecular Biosystems*, 6(2):289–304, 2010.

[44] I. Dinu, J. D. Potter, T. Mueller, Q. Liu, A. J. Adewale, G. S. Jhangri, G. Einecke, K. S. Famulski, P. Halloran, and Y. Yasui. Improving gene set analysis of microarray data by sam-gs. *BMC*, 8:242, 2007.

[45] S. R. Eddy. What is bayesian statistics? *Nature Biotechnology*, 22(9):1177–1178, 2004.

[46] J. S. Edwards and B. O. Palsson. The escherichia coli mg1655 in silico metabolic genotype: its definition, characteristics, and capabilities. *Proc Natl Acad Sci U S A*, 97(10):5528–5533, May 2000.

[47] P. H. C. Eilers. Deconvolution of spike trains using an l0 penalty. In J.G. Booth, editor, *Proceedings of the 24th International Workshop on Statistical Modelling, Ithaca 20-24 July*, pages 130–137. Ithaca, NY: Cornell University, Department of Biological Statistics and Computational Biology, 2009.

[48] F. Emmert-Streib and G. V. Glazko. Pathway analysis of expression data: Deciphering functional building blocks of complex diseases. *Plos Computational Biology*, 7(5):e1002053, 2011.

[49] H. W. Engl, C. Flamm, P. Kugler, J. Lu, S. Muller, and P. Schuster. Inverse problems in systems biology. *Inverse Problems*, 25:123014, 2009.

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

[51] X. Feng, Y. Xu, Y. Chen, and Y. J. Tang. Integrating flux balance analysis into kinetic models to decipher the dynamic metabolism of shewanella oneidensis mr-1. *PLoS Comput Biol*, 8(2):e1002376, Feb 2012.

[52] T. Ferguson. Linear programming: A concise introduction. (http://www.math.ucla.edu/ tom/LP.pdf), 2011.

[53] O. Fiehn. Combining genomics, metabolome analysis, and biochemical modelling to understand metabolic networks. *Comp Funct Genomics*, 2(3):155–168, 2001.

[54] R.A. Fisher. On the "probable error" of a coefficient of correlation deduced from a small sample. *Metron*, 1(4):3–32, 1921.

[55] J. Forster, I. Famili, P. Fu, B. O. Palsson, and J. Nielsen. Genome-scale reconstruction of the saccharomyces cerevisiae metabolic network. *Genome Research*, 13(2):244–253, 2003.

[56] O. Frick and C. Wittmann. Characterization of the metabolic shift between oxidative and fermentative growth in saccharomyces cerevisiae by comparative c-13 flux analysis. *Microbial Cell Factories*, 4:30, 2005.

[57] N. Friedman, M. Linial, I. Nachman, and D. Pe'er. Using bayesian networks to analyze expression data. *J Comput Biol*, 7(3-4):601–620, 2000.

[58] M. L. Giuseppin and N. A. van Riel. Metabolic modeling of saccharomyces cerevisiae using the optimal control of homeostasis: a cybernetic model definition. *Metab Eng*, 2(1):14–33, Jan 2000.

[59] J. J. Goeman, S. A. van de Geer, F. de Kort, and H. C. van Houwelingen. A global test for groups of genes: testing association with a clinical outcome. *Bioinformatics*, 20(1):93–99, 2004.

[60] J. J. Goeman, S. A. van de Geer, and H. C. van Houwelingen. Testing against a high dimensional alternative. *Journal of the*

Royal Statistical Society Series B-Statistical Methodology, 68:477–493, 2006.

[61] I. Golding. Decision making in living cells: lessons from a simple system. *Annu Rev Biophys*, 40:63–80, 2011.

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

[63] H. Hache, H. Lehrach, and R. Herwig. Reverse engineering of gene regulatory networks: A comparative study. *EURASIP J Bioinform Syst Biol*, 2009:617281, 2009.

[64] T. Hastie, R. Tibshirami, and J. Friedman. *The elements of statistical learning. Data mining, inference and prediction.* Springer-Verlag, New York, 2001.

[65] B. D. Heavner, K. Smallbone, B. Barker, P. Mendes, and L. P. Walker. Yeast 5 - an expanded reconstruction of the saccharomyces cerevisiae metabolic network. *BMC Systems Biology*, 6:55, 2012.

[66] D. M. Hendrickx, M. M. W. B. Hendriks, P. H. C. Eilers, A. K. Smilde, and H. C. J. Hoefsloot. Reverse engineering of metabolic networks, a critical assessment. *Molecular Biosystems*, 7(2):511–520, 2011.

[67] D. M. Hendrickx, H. C. J. Hoefsloot, M. M. W. B. Hendriks, A. B. Canelas, and A. K. Smilde. Global test for metabolic pathway differences between conditions. *Anal Chim Acta*, 719:8–15, Mar 2012.

[68] J. L. Hjersted and M. A. Henson. Steady-state and dynamic flux balance analysis of ethanol production by saccharomyces cerevisiae. *IET Syst Biol*, 3(3):167–179, May 2009.

[69] K. Hollywood, D. R. Brison, and R. Goodacre. Metabolomics: Current technologies and future trends. *Proteomics*, 6(17):4716–4723, 2006.

[70] H. G. Holzhütter. The principle of flux minimization and its application to estimate stationary fluxes in metabolic networks. *Eur J Biochem*, 271(14):2905–2922, Jul 2004.

[71] T. Ideker, T. Galitski, and L. Hood. A new approach to decoding life: Systems biology. *Annual Review of Genomics and Human Genetics*, 2:343–372, 2001.

[72] T. Ideker and N. J. Krogan. Differential network biology. *Mol Syst Biol*, 8:565, 2012.

[73] F. Jacob and J. Monod. Genetic regulatory mechanisms in the synthesis of proteins. *J Mol Biol*, 3:318–356, Jun 1961.

[74] N. Jamshidi and B. O. Palsson. Top-down analysis of temporal hierarchy in biochemical reaction networks. *PLoS Comput Biol*, 4(9):e1000177, 2008.

[75] J. J. Jansen, E. Szymanska, H. C. J. Hoefsloot, D. M. Jacobs, K. Strassburg, and A. K. Smilde. Between metabolite relationships: an essential aspect of metabolic change. *Metabolomics*, 8(3):422–432, Jun 2012.

[76] Z. Jiang and R. Gentleman. Extensions to gene set enrichment. *Bioinformatics*, 23(3):306–313, 2007.

[77] I. T. Jolliffe. *Principal component analysis*. Springer-Ver, 1986.

[78] P. Jouhten, M. Wiebe, and M. Penttila. Dynamic flux balance analysis of the metabolism of saccharomyces cerevisiae during the shift from fully respirative or respirofermentative metabolic states to anaerobiosis. *FEBS J*, 279(18):3338–3354, Sep 2012.

[79] M. S. Jurica, A. Mesecar, P. J. Heath, S. Wuxian, T. Nowak, and B. L. Stoddard. The allosteric regulation of pyruvate kinase by fructose-1,6-bisphosphate. *Structure*, 6(2):195–210, 1998.

[80] P. Kahlem, A. DiCara, M. Durot, J. M. Hancock, E. Klipp, V. Schchter, E. Segal, I. Xenarios, E. Birney, and L. Mendoza. Strengths and weaknesses of selected modeling methods used in systems biology. http://tainguyenso.vnu.edu.vn/jspui/handle/123456789/16745, Nov 2011.

[81] T. Kamminga. Short-term dynamics of glycolysis in saccharomyces cerevisiae expressing arginine kinase. Master's thesis, Department of Biotechnology, Delft University of Technology, February, 2007.

[82] M. Kanehisa and S. Goto. Kegg: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Research*, 28(1):27–30, 2000.

[83] M. Kanehisa, S. Goto, M. Furumichi, M. Tanabe, and M. Hirakawa. Kegg for representation and analysis of molecular networks involving diseases and drugs. *Nucleic Acids Research*, 38:D355–D360, 2010.

[84] M. Kanehisa, S. Goto, M. Hattori, K. F. Aoki-Kinoshita, M. Itoh, S. Kawashima, T. Katayama, M. Araki, and M. Hirakawa. From genomics to chemical genomics: new developments in kegg. *Nucleic Acids Research*, 34:D354–D357, 2006.

[85] M. Kankainen, P. Gopalacharyulu, L. Holm, and M. Oresic. Mpea - metabolite pathway enrichment analysis. *Bioinformatics*, 27(13):1878–9, 2011.

[86] D. M. Keenan and J. D. Veldhuis. Divergent gonadotropin-gonadal dose-responsive coupling in healthy young and aging men. *Am J Physiol Regul Integr Comp Physiol*, 286(2):R 381–9, 2004.

[87] B. Kholodenko, M. B. Yaffe, and W. Kolch. Computational approaches for analyzing information flow in biological networks. *Sci Signal*, 5(220):re1, Apr 2012.

[88] J. Kim, D. G. Bates, I. Postlethwaite, P. Heslop-Harrison, and K. H. Cho. Linear time-varying models can reveal non-linear in-

teractions of biomolecular regulatory networks using multiple time-series data. *Bioinformatics*, 24(10):1286–1292, 2008.

[89] S. J. Kim and D. J. Volsky. Page: Parametric analysis of gene set enrichment. *BMC Bioinformatics*, 6:144, 2005.

[90] H. Kitano. Biological robustness. *Nat Rev Genet*, 5(11):826–837, Nov 2004.

[91] S. Kleessen and Z. Nikoloski. Dynamic regulatory on/off minimization for biological systems under internal temporal perturbations. *BMC Syst Biol*, 6(1):16, 2012.

[92] C. Klein, A. Marino, M.-F. Sagot, P. V. Paulo Milreu, and M. Brilli. Structural and dynamical analysis of biological networks. *Brief Funct Genomics*, Aug 2012.

[93] E. Klipp, R. Herwig, A. Kowald, C. Wierling, and H. Lehrach. *Systems Biology in Practice. Concepts, Implementation and Application.* Wiley-VCH Verlag GmbH & Co. KgaA, Weinheim, Germany., 2005.

[94] M. M. Koek, B. Muilwijk, M. J. van der Werf, and T. Hankemeier. Microbial metabolomics with gas chromatography/mass spectrometry. *Analytical Chemistry*, 78(4):1272–1281, 2006.

[95] P. Kohl, E. J. Crampin, T. A. Quinn, and D. Noble. Systems biology: an approach. *Clin Pharmacol Ther*, 88(1):25–33, Jul 2010.

[96] P. Kok, F. Roelfsema, M. Frolich, J. van Pelt, A. E. Meinders, and H. Pijl. Activation of dopamine d2 receptors lowers circadian leptin concentrations in obese women. *J Clin Endocrinol Metab*, 91(8):3236–40, 2006.

[97] P. Kok, F. Roelfsema, M. Frolich, J. van Pelt, M. P. Stokkel, A. E. Meinders, and H. Pijl. Activation of dopamine d2 receptors simultaneously ameliorates various metabolic features of obese women. *Am J Physiol Endocrinol Metab*, 291(5):E 1038–43, 2006.

[98] S. W. Kok, A. E. Meinders, S. Overeem, G. J. Lammers, F. Roelfsema, M. Frolich, and H. Pijl. Reduction of plasma leptin levels and loss of its circadian rhythmicity in hypocretin (orexin)-deficient narcoleptic humans. *Journal of Clinical Endocrinology and Metabolism*, 87(2):805–809, 2002.

[99] S. W. Kok, F. Roelfsema, S. Overeem, G. J. Lammers, M. Frolich, A. E. Meinders, and H. Pijl. Pulsatile lh release is diminished, whereas fsh secretion is normal, in hypocretin-deficient narcoleptic men. *American Journal of Physiology-Endocrinology and Metabolism*, 287(4):E 630–E 636, 2004.

[100] S. W. Kok, F. Roelfsema, S. Overeem, G. J. Lammers, M. Frolich, A. E. Meinders, and H. Pijl. Altered setting of the pituitary-thyroid ensemble in hypocretin-deficient narcoleptic men. *Am J Physiol Endocrinol Metab*, 288(5):E 892–9, 2005.

[101] S. W. Kok, F. Roelfsema, S. Overeem, G. J. Lammers, R. L. Strijers, M. Frolich, A. E. Meinders, and H. Pijl. Dynamics of the pituitary-adrenal ensemble in hypocretin-deficient narcoleptic humans: blunted basal adrenocorticotropin release and evidence for normal time-keeping by the master pacemaker. *J Clin Endocrinol Metab*, 87(11):5085–91, 2002.

[102] S. W. Kong, W. T. Pu, and P. J. Park. A multivariate approach for integrating genome-wide expression data and biological knowledge. *Bioinformatics*, 22(19):2373–2380, 2006.

[103] O. Kotte, J. B. Zaugg, and M. Heinemann. Bacterial adaptation through distributed sensing of metabolic fluxes. *Molecular Systems Biology*, 6:355, 2010.

[104] M. T A P Kresnowati, W. A. van Winden, M. J H Almering, A. ten Pierick, C. Ras, T. A. Knijnenburg, P. Daran-Lapujade, J. T. Pronk, J. J. Heijnen, and J. M. Daran. When transcriptome meets metabolome: fast cellular responses of yeast to sudden relief of glucose limitation. *Mol Syst Biol*, 2:49, 2006.

[105] A. Kuchina, L. Espinar, J. Garcia-Ojalvo, and G. M. Sel. Reversible and noisy progression towards a commitment point enables adaptable and reliable cellular decision-making. *PLoS Comput Biol*, 7(11):e1002273, Nov 2011.

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

[107] P. Lecca and C. Priami. Biological network inference for drug discovery. *Drug Discov Today*, page DOI 10.1016/j.drudis.2012.11.001, Nov 2012.

[108] S. Lecessie and H. C. Vanhouwelingen. Testing the fit of a regression-model via score tests in random effects models. *Biometrics*, 51(2):600–614, 1995.

[109] J. M. Lee, E. P. Gianchandani, and J. A. Papin. Flux balance analysis in the era of metabolomics. *Brief Bioinform*, 7(2):140–150, 2006.

[110] R. W. Leighty and M. R. Antoniewicz. Dynamic metabolic flux analysis (dmfa): a framework for determining fluxes at metabolic non-steady state. *Metab Eng*, 13(6):745–755, Nov 2011.

[111] E. Libby, T. J. Perkins, and P. S. Swain. Noisy information processing through transcriptional regulation. *Proc Natl Acad Sci U S A*, 104(17):7151–7156, Apr 2007.

[112] F. Llaneras and J. Pico. Stoichiometric modelling of cell metabolism. *J Biosci Bioeng*, 105(1):1–11, Jan 2008.

[113] O. H. Lowry, J. Carter, J. B. Ward, and L. Glaser. The effect of carbon and nitrogen sources on the level of metabolic intermediates in escherichia coli. *J Biol Chem*, 246(21):6511–21, 1971.

[114] D. G. Luenberger. *Introduction to linear and nonlinear programming*. Addison-Wesley, 1973.

[115] J. Lundgren. Splinefit. (http://www.mathworks.com/matlabcentral /fileexchange/13812-splinefit),Retrieved September 26, 2011, 2007.

[116] R. Y. Luo, S. Liao, G. Y. Tao, Y. Y. Li, S. Zeng, Y. X. Li, and Q. Luo. Dynamic analysis of optimality in myocardial energy metabolism under normal and ischemic conditions. *Mol Syst Biol*, 2:2006.0031, 2006.

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

[118] D. Machado, R. S. Costa, E. C. Ferreira, I. Rocha, and B. Tidor. Exploring the gap between dynamic and constraint-based models of metabolism. *Metab Eng*, 14(2):112–119, Mar 2012.

[119] D. Machado, R. S. Costa, M. Rocha, E. C. Ferreira, B. Tidor, and I. Rocha. Modeling formalisms in systems biology. *AMB Express*, 1:45, 2011.

[120] R. Mahadevan, J. S. Edwards, and F. J. Doyle. Dynamic flux balance analysis of diauxic growth in escherichia coli. *Biophys J*, 83(3):1331–1340, Sep 2002.

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

[122] A. M. Mahmood, M. R. Kuppa, and K. K. Reddi. A new decision tree induction using composite splitting criterion. *Journal of Applied Computer Science & Mathematics*, 9(4):67–71, 2010.

[123] A. M. Martins, D. Camacho, J. Shuman, W. Sha, P. Mendes, and V. Shulaev. A systems biology study of two distinct growth phases

of saccharomyces cerevisiae cultures. *Current Genomics*, 5(8):649–663, 2004.

[124] M. R. Mashego, W. M. Gulik, and J. J. Heijnen. Metabolome dynamic responses of saccharomyces cerevisiae to simultaneous rapid perturbations in external electron acceptor and electron donor. *Fems Yeast Research*, 7(1):48–66, 2007.

[125] M. R. Mashego, W. M. van Gulik, J. L. Vinke, D. Visser, and J. J. Heijnen. In vivo kinetics with rapid perturbation experiments in saccharomyces cerevisiae using a second-generation bioscope. *Metabolic Engineering*, 8(4):370–383, 2006.

[126] C. K. Mathews and K. E. van Holde. *Biochemistry*. The Benjamin/Cummings Publishing Company, Inc. Menlo Park, California, 1996.

[127] MATLAB®. Version 7.5.0, copyright©, 1984-2007,The Mathworks Inc.

[128] MATLAB®. Version 7.11.0.(r2010b), microsoft windows xp version 5.1., copyright©, 1984-2010,The Mathworks Inc.

[129] Y. Matsuoka and K. Shimizu. The relationships between the metabolic fluxes ans 13c-labeled isotopomer distribution for the flux analysis of the main metabolic pathways. *Biochemical Engineering Journal*, 49:326–336, 2010.

[130] P. McCullagh and J. A. Nelder. *Generalized Linear Models*. Chapman and Hall, Boca Raton, USA, 1989.

[131] P. Mendes, D. Camacho, and A. de la Fuente. Modelling and simulation for metabolomics data analysis. *Biochem Soc Trans*, 33(6):1427–1429, 2005.

[132] D. Molenaar, R. van Berlo, D. de Ridder, and B. Teusink. Shifts in growth strategies reflect tradeoffs in cellular economics. *Mol Syst Biol*, 5:323, 2009.

[133] V. K. Mootha, C. M. Lindgren, K. F. Eriksson, A. Subramanian, S. Sihag, J. Lehar, P. Puigserver, E. Carlsson, M. Ridderstrale, E. Laurila, N. Houstis, M. J. Daly, N. Patterson, J. P. Mesirov, T. R. Golub, P. Tamayo, B. Spiegelman, E. S. Lander, J. N. Hirschhorn, D. Altshuler, and L. C. Groop. Pgc-1 alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nature Genetics*, 34(3):267–273, 2003.

[134] K. Morgenthal, W. Weckwerth, and R. Steuer. Metabolomic networks in plants: Transitions from pattern recognition to biological interpretation. *Biosystems*, 83(2-3):108–117, 2006.

[135] M. Muller-Linow, W. Weckwerth, and M. T. Hutt. Consistency analysis of metabolic correlation networks. *BMC Syst Biol*, 1:44, 2007.

[136] D. Nagrath, M. Avila-Elchiver, F. Berthiaume, A. W. Tilles, A. Messac, and M. L. Yarmush. Integrated energy and flux balance based multiobjective framework for large-scale metabolic networks. *Ann Biomed Eng*, 35(6):863–885, Jun 2007.

[137] D. Nam and S. Y. Kim. Gene-set approach for expression pattern analysis. *Briefings in Bioinformatics*, 9(3):189–197, 2008.

[138] J. K. Nicholson, J. Connelly, J. C. Lindon, and E. Holmes. Metabonomics: a platform for studying drug toxicity and gene function. *Nature Reviews Drug Discovery*, 1:153–161, 2002.

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

[140] I. E. Nikerel, A. B. Canelas, S. J. Jol, P. J. T. Verheijen, and J. J. Heijnen. Construction of kinetic models for metabolic reaction networks: Lessons learned in analysing short-tern stimulus response data. *Mathematical and Computer Modelling of Dynamical Systems*, 17(3):243–260, 2011.

[141] A. Nilsson, I. L. Pahlman, P. A. Jovall, A. Blomberg, C. Larsson, and L. Gustafsson. The catabolic capacity of saccharomyces cerevisiae is preserved to a higher extent during carbon compared to nitrogen starvation. *Yeast*, 18(15):1371–1381, 2001.

[142] T. L. Nissen, U. Schulze, J. Nielsen, and J. Villadsen. Flux distributions in anaerobic, glucose-limited continuous cultures of saccharomyces cerevisiae. *Microbiology-Uk*, 143:203–218, 1997.

[143] S. Noack, K. Noh, M. Moch, M. Marco Oldiges, and W. Wiechert. Stationary versus non-stationary (13)c-mfa: a comparison using a consistent dataset. *J Biotechnol*, 154(2-3):179–190, Jul 2011.

[144] J. O'Grady, J. Schwender, Y. Shachar-Hill, and J. A. Morgan. Metabolic cartography: experimental quantification of metabolic fluxes from isotopic labelling studies. *J Exp Bot*, 63(6):2293–2308, Mar 2012.

[145] M. Oldiges, S. Lutz, S. Pflug, K. Schroer, N. Stein, and C. Wiendahl. Metabolomics: current state and evolving methodologies and tools. *Appl Microbiol Biotechnol*, 76(3):495–511, Sep 2007.

[146] B. G. Olivier and J. L. Snoep. Web-based kinetic modelling using jws online. *Bioinformatics*, 20(13):2143–2144, 2004.

[147] J. D. Orth, I. Thiele, and B. O. Palsson. What is flux balance analysis? *Nat Biotechnol*, 28(3):245–248, Mar 2010.

[148] S. Overeem, S. W. Kok, G. J. Lammers, A. A. Vein, M. Frolich, A. E. Meinders, F. Roelfsema, and H. Pijl. Somatotropic axis in hypocretin-deficient narcoleptic humans: altered circadian distribution of gh-secretory events. *Am J Physiol Endocrinol Metab*, 284(3):E 641–7, 2003.

[149] A. Paldi. What makes the cell differentiate? *Prog Biophys Mol Biol*, page doi:10.1016/j.pbiomolbio.2012.04.003, Apr 2012.

[150] B. O. Palsson. *Systems Biology: Properties of Reconstructed Networks [Kindle Edition]*. Cambridge University Press, 2006.

[151] K. Pearson. Mathematical contributions to the theory of evolution. iii. regression, heredity, and panmixia. *Philosophical Transactions of the Royal Society a-Mathematical Physical and Engineering Sciences*, 187:253–318, 1896.

[152] T. J. Perkins and P. S. Swain. Strategies for cellular decision-making. *Mol Syst Biol*, 5:326, 2009.

[153] J. Piskur, E. Rozpedowska, S. Polakova, A. Merico, and C. Compagno. How did saccharomyces evolve to become a good brewer? *Trends Genet*, 22(4):183–186, Apr 2006.

[154] J. Postmus, A. B. Canelas, J. Bouwman, B. M. Bakker, W. van Gulik, M. J. T. de Mattos, S. Brul, and G. J. Smits. Quantitative analysis of the high temperature-induced glycolytic flux increase in saccharomyces cerevisiae reveals dominant metabolic regulation. *Journal of Biological Chemistry*, 283(35):23524–23532, 2008.

[155] C. F. Quo, C. Kaddi, J. H. Phan, A. Zollanvari, M. Xu, M. D. Wang, and G. Alterovitz. Reverse engineering biomolecular systems using -omic data: challenges, progress and opportunities. *Brief Bioinform*, 13(4):430–445, Jul 2012.

[156] R. Ramakrishna, J. S. Edwards, A. McCulloch, and B. O. Palsson. Flux-balance analysis of mitochondrial energy metabolism: consequences of systemic stoichiometric constraints. *Am J Physiol Regul Integr Comp Physiol*, 280(3):R695–R704, Mar 2001.

[157] C. R. Rao. Score test: Historical review and recent developments. *Statistics for Industry and Technology*, Part I:3–20, 2005.

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

[159] J. L. Reed. Shrinking the metabolic solution space using experimental datasets. *PLoS Comput Biol*, 8(8):e1002662, Aug 2012.

[160] D. M. Rocke and S. Lorenzato. A two-component model for measurement error in analytical chemistry. *Technometrics*, 37(2):176–184, 1995.

[161] F. Rodrigues, P. Ludovico, and C. Leao. Sugar metabolism in yeasts : an overview of aerobic and anaerobic glucose catabolism. In *In ROSA, Carlos ; PETER, Gabor, ed. lit. Biodiversity and ecophysiology of yeasts.*, pages 101–121. Springer, Berlin, 2006.

[162] J. M. Rohwer. Kinetic modelling of plant metabolic pathways. *J Exp Bot*, 63(6):2275–2292, Mar 2012.

[163] S. Rossell, C. C. van der Weijden, A. Lindenbergh, A. van Tuijl, C. Francke, B. M. Bakker, and H. V. Westerhoff. Unraveling the complexity of flux regulation: A new method demonstrated for nutrient starvation in saccharomyces cerevisiae. *Proceedings of the National Academy of Sciences of the United States of America*, 103(7):2166–2171, 2006.

[164] C. M. Rubingh, S. Bijlsma, R. H. Jellema, K. M. Overkamp, M. J. van der Werf, and A. K. Smilde. Analyzing longitudinal microbial metabolomics data. *J Proteome Res*, 8(9):4319–27, 2007.

[165] E. Saccenti, J. A. Westerhuis, A. K. Smilde, M. J. van der Werf, J. A. Hageman, and M. M. W. B. Hendriks. Simplivariate models: Uncovering the underlying biology in functional genomics data. *PLOS one*, 6(6):e20747, 2011.

[166] M. Samoilov, A. Arkin, and J. Ross. On the deduction of chemical reaction pathways from measurements of time series of concentrations. *Chaos*, 11(1):108–114, 2001.

[167] D. Sarkar, S. Falcon, and R. Gentleman. Using categories defined by chromosome bands. bioconductor, version 2.5, http://www.bioconductor.org, August 2008.

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

J. J. Tyson. Challenges for modeling and simulation methods in systems biology. In *Proceedings of the 2006 Winter Simulation Conference*, 2006.

[169] U. Schaefer, W. Boos, R. Takors, and D. Weuster-Botz. Automated sampling device for monitoring intracellular metabolite dynamics. *Analytical Biochemistry*, 270(1):88–96, 1999.

[170] J. Schäfer and K. Strimmer. An empirical bayes approach to inferring large-scale gene association networks. *Bioinformatics*, 21(6):754–764, 2005.

[171] J. Schaub, C. Schiesling, M. Reuss, and M. Dauner. Integrated sampling procedure for metabolome analysis. *Biotechnology Progress*, 22(5):1434–1442, 2006.

[172] H. Schmidt, K. H. Cho, and E. W. Jacobsen. Identification of small scale biochemical networks based on general type system perturbations. *Febs Journal*, 272(9):2141–2151, 2005.

[173] M. Schmidt. Least squares optimization with l1-norm regularization., 2005.

[174] J. R. Schott. *Matrix analysis for statistics. Second edition.* Wiley Series in Probability and Statistics. Wiley-Interscience, Hoboken, New Jersey, USA, 2005.

[175] R. Schuetz, L. Kuepfer, and U. Sauer. Systematic evaluation of objective functions for predicting intracellular fluxes in escherichia coli. *Mol Syst Biol*, 3:119, 2007.

[176] R. Schuetz, N. Zamboni, M. Zampieri, M. Heinemann, and U. Sauer. Multidimensional optimality of microbial metabolism. *Science*, 336(6081):601–604, 2012.

[177] A. Schwabe, M. Dobrzyski, K. Rybakova, P. Verschure, and F. J. Bruggeman. *Origins of Stochastic Intracellular Processes and Consequences for Cell-to-Cell Variability and Cellular Survival Strategies.*, volume 500 of *Methods in Enzymology*, chapter twenty-eight, pages 597–625. Burlington: Academic Press, 2011.

[178] D. Segré, D. Vitkup, and G. M. Church. Analysis of optimality in natural and perturbed metabolic networks. *Proc Natl Acad Sci U S A*, 99(23):15112–15117, 2002.

[179] J. Shi and M. G. Walker. Gene set enrichment analysis (gsea) for interpreting gene expression profiles. *Current Bioinformatics*, 2(2):133–137, 2007.

[180] E. S. Snitkin and D. Segré. Optimality criteria for the prediction of metabolic fluxes in yeast mutants. *Genome Inform*, 20:123–134, 2008.

[181] J. L. Snoep, F. J. Bruggeman, B. G. Olivier, and H. V. Westerhoff. Towards building the silicon cell: a modular approach. *Biosystems*, 83(2-3):207–16, 2006.

[182] E. Sontag, A. Kiyatkin, and B. N. Kholodenko. Inferring dynamic architecture of cellular networks using time series of gene expression, protein and metabolite data. *Bioinformatics*, 20(12):1877–1886, Aug 2004.

[183] A. Spanos. *Probability theory and statistical inference. Econometric modeling with observational data*. Cambridge University Press, Cambridge, UK, 1999.

[184] J. Srividhya, M. A. Mourao, E. J. Crampin, and S. Schnell. Enzyme catalyzed reactions: From experiment to computational mechanism reconstruction. *Computational Biology and Chemistry*, 34(1):11–18, 2010.

[185] H. Starling. The croonian lectures on the chemical correlation of the functions of the body. *Lancet*, II:339–341, 1905.

[186] M. Steinfath, D. Repsilber, M. Scholz, D. Walther, and J. Selbig. Integrated data analysis for genome-wide research. *EXS*, 97:309–329, 2007.

[187] J. Stelling. Mathematical models in microbial systems biology. *Curr Opin Microbiol*, 7(5):513–518, Oct 2004.

[188] J. Stelling, U. Sauer, Z. Szallasi, F. J. Doyle, and J. Doyle. Robustness of cellular functions. *Cell*, 118(6):675–685, Sep 2004.

[189] R. Steuer. Review: on the analysis and interpretation of correlations in metabolomic data. *Brief Bioinform*, 7(2):151–158, 2006.

[190] R. Steuer. Computational approaches to the topology, stability and dynamics of metabolic networks. *Phytochemistry*, 68(16-18):2139–2151, 2007.

[191] R. Steuer, T. Gross, J. Selbig, and B. Blasius. Structural kinetic modeling of metabolic networks. *Proc Natl Acad Sci U S A*, 103(32):11868–11873, 2006.

[192] R. Steuer, J. Kurths, O. Fiehn, and W. Weckwerth. Observing and interpreting correlations in metabolomic networks. *Bioinformatics*, 19(8):1019–1026, 2003.

[193] IBM®ILOG®CPLEX® Optimization Studio. V12.2, microsoft windows xp version 5.1., copyright©, 1987-2010,IBM Corp.

[194] A. Subramanian, P. Tamayo, V. K. Mootha, S. Mukherjee, B. L. Ebert, M. A. Gillette, A. Paulovich, S. L. Pomeroy, T. R. Golub, E. S. Lander, and J. P. Mesirov. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A*, 102(43):15545–50, 2005.

[195] M. Swat, S. M. Kielbasa, S. Polak, B. Olivier, F. J. Bruggeman, M. Q. Tulloch, J. L. Snoep, A. J. Verhoeven, and H. V. Westerhoff. What it takes to understand and cure a living system: computational systems biology and a systems biology-driven pharmacokinetics-pharmacodynamics platform. *Interface Focus*, 1(1):16–23, Feb 2011.

[196] Z. Szallasi, J. Stelling, and V. Periwal(eds.). *System Modeling in Cellular Biology: From Concepts to Nuts and Bolts*. MIT Press, Massachusetts, 2006.

[197] Y. J. Tang, J. S. Hwang, D. E. Wemmer, and J. D. Keasling. Shewanella oneidensis mr-1 fluxome under various oxygen conditions. *Appl Environ Microbiol*, 73(3):718–729, Feb 2007.

[198] C. F. Taylor, D. Field, S.-A. Sansone, J. Aerts, R. Apweiler, M. Ashburner, C. A. Ball, P.-A. Binz, M. Bogue, T. Booth, A. Brazma, R. R. Brinkman, A. M. Clark, E. W. Deutsch, O. Fiehn, J. Fostel, P. Ghazal, F. Gibson, T. Gray, G. Grimes, J. M. Hancock, N. W. Hardy, H. Hermjakob, R. K. Julian, M. Kane, C. Kettner, C. Kinsinger, E. Kolker, M. Kuiper, N. Le Novere, J. Leebens-Mack, S. E. Lewis, P. Lord, A.-M. Mallon, N. Marthandan, H. Masuya, R. McNally, A. Mehrle, N. Morrison, S. Orchard, J. Quackenbush, J. M. Reecy, D. G. Robertson, P. Rocca-Serra, H. Rodriguez, H. Rosenfelder, J. Santoyo-Lopez, R. H. Scheuermann, D. Schober, B. Smith, J. Snape, C. J. Stoeckert, K. Tipton, P. Sterk, A. Untergasser, J. Vandesompele, and S. Wiemann. Promoting coherent minimum reporting guidelines for biological and biomedical investigations: the mibbi project. *Nat Biotechnol*, 26(8):889–896, Aug 2008.

[199] H. Taymaz-Nikerel, M. de Mey, C. Ras, A. ten Pierick, R. M. Seifar, J. C. Van Dam, J. J. Heijnen, and W. M. Van Glilik. Development and application of a differential method for reliable metabolome analysis in escherichia coli. *Analytical Biochemistry*, 386(1):9–19, 2009.

[200] N. Tenazinha and S. Vinga. A survey on methods for modeling and analyzing integrated biological networks. *IEEE/ACM Trans Comput Biol Bioinform*, 8(4):943–958, 2011.

[201] B. Teusink, J. Passarge, C. A. Reijenga, E. Esgalhado, C. C. van der Weijden, M. Schepper, M. C. Walsh, B. M. Bakker, K. van Dam, H. V. Westerhoff, and J. L. Snoep. Can yeast glycolysis be understood in terms of in vitro kinetics of the constituent enzymes? testing biochemistry. *European Journal of Biochemistry*, 267(17):5313–5329, 2000.

[202] U. Theobald, W. Mailinger, M. Baltes, M. Rizzi, and M. Reuss. In vivo analysis of metabolic dynamics in saccharomyces cerevisiae .1. experimental observations. *Biotechnology and Bioengineering*, 55(2):305–316, 1997.

[203] U. Theobald, W. Mailinger, M. Reuss, and M. Rizzi. In-vivo analysis of glucose-induced fast changes in yeast adenine-nucleotide pool applying a rapid sampling technique. *Analytical Biochemistry*, 214(1):31–37, 1993.

[204] C. A. Tsai and J. J. Chen. Multivariate analysis of variance test for gene set analysis. *Bioinformatics*, 25(7):897–903, 2009.

[205] R. Ursem, Y. Tikunov, A. Bovy, R. van Berloo, and F. van Eeuwijk. A correlation network approach to metabolic data analysis for tomato fruits. *Euphytica*, 161(1-2):181–193, 2008.

[206] M. F. Van Batenburg, L. Coulier, F. van Eeuwijk, A. K. Smilde, and J. A. Westerhuis. New figures of merit for comprehensive functional genomics data: the metabolomics case. *Anal Chem*, 83(9):3267–3274, May 2011.

[207] R. A. van den Berg, H. C. J. Hoefsloot, J. A. Westerhuis, A. K. Smilde, and M. J. van der Werf. Centering, scaling, and transformations: improving the biological information content of metabolomics data. *BMC Genomics*, 7:142, 2006.

[208] M. J. van der Werf, K. M. Overkamp, B. Muilwijk, L. Coulier, and T. Hankemeier. Microbial metabolomics: Toward a platform with full metabolome coverage. *Analytical Biochemistry*, 370(1):17–25, 2007.

[209] K. van Eunen. *The multifarious and dynamic regulation of the living cell.* PhD thesis, Free University of Amsterdam, 2010.

[210] W. M. van Gulik. Fast sampling for quantitative microbial metabolomics. *Curr Opin Biotechnol*, 21(1):27–34, 2010.

[211] N. A. W. van Riel. Dynamic modelling and analysis of biochemical networks: mechanism-based models and model-based experiments. *Brief Bioinform*, 7(4):364–374, Dec 2006.

[212] W. A. van Winden, J. C. van Dam, C. Ras, R. J. Kleijn, J. L. Vinke, W. M. van Gulik, and J. J. Heijnen. Metabolic-flux analysis of saccharomyces cerevisiae cen.pk113-7d based on mass isotopomer measurements of (13)c-labeled primary metabolites. *FEMS Yeast Res*, 5(6-7):559–568, Apr 2005.

[213] W. Vance, A. Arkin, and J. Ross. Determination of causal connectivities of species in reaction networks. *Proc Natl Acad Sci U S A*, 99(9):5816–5821, 2002.

[214] A. Varma and B. O. Palsson. Metabolic flux balancing - basic concepts, scientific and practical use. *Bio-Technology*, 12(10):994–998, 1994.

[215] S. Vaseghi, A. Baumeister, M. Rizzi, and M. Reuss. In vivo dynamics of the pentose phosphate pathway in saccharomyces cerevisiae. *Metab Eng*, 1(2):128–40, 1999.

[216] D. Visser, G. A. van Zuylen, J. C. van Dam, M. R. Eman, A. Proll, C. Ras, L. Wu, W. M. van Gulik, and J. J. Heijnen. Analysis of in vivo kinetics of glycolysis in aerobic saccharomyces cerevisiae by application of glucose and ethanol pulses. *Biotechnology and Bioengineering*, 88(2):157–167, 2004.

[217] D. Visser, G. A. van Zuylen, J. C. van Dam, A. Oudshoorn, M. R. Eman, C. Ras, W. M. van Gulik, J. Frank, G. W. K. van Dedem, and J. J. Heijnen. Rapid sampling for analysis of in vivo kinetics using the bioscope: A system for continuous-pulse experiments. *Biotechnology and Bioengineering*, 79(6):674–681, 2002.

[218] G. A. Viswanathan, J. Seto, S. Patil, G. Nudelman, and S. C. Sealfon. Getting started in biological pathway construction and analysis. *PLoS Comput Biol*, 4(2):e16, Feb 2008.

[219] T. D. Vo, H. J. Greenberg, and B. O. Palsson. Reconstruction and functional characterization of the human mitochondrial metabolic network based on proteomic and biochemical data. *J Biol Chem*, 279(38):39532–39540, Sep 2004.

[220] S. A. Wahl, K. Noh, and W. Wiechert. 13c labeling experiments at metabolic nonstationary conditions: an exploratory study. *BMC Bioinformatics*, 9:152, 2008.

[221] G.M. Walker. *Yeast Physiology and Biotechnology.* John Wiley & Sons. Chichester, West Sussex, England, 1998.

[222] Q.-Z. Wang, C.-Y. Wu, T. Chen, X. Chen, and X.-M. Zhao. Integrating metabolomics into a systems biology framework to exploit metabolic complexity: strategies and applications in microorganisms. *Appl Microbiol Biotechnol*, 70(2):151–161, Mar 2006.

[223] W. Weckwerth and O. Fiehn. Can we discover novel pathways using metabolomic analysis? *Current Opinion in Biotechnology*, 13(2):156–160, 2002.

[224] W. Weckwerth and K. Morgenthal. Metabolomics: from pattern recognition to biological interpretation. *Drug Discov Today*, 10(22):1551–1558, Nov 2005.

[225] J. A. Westerhuis, E. P. P. A. Derks, H. C. J. Hoefsloot, and A. K. Smilde. Grey component analysis. *Journal of Chemometrics*, 21(10-11):474–485, 2007.

[226] J. A. Westerhuis, H. C. J. Hoefsloot, S. Smit, D. J. Vis, A. K. Smilde, E. J. J. van Velzen, J. P. M. van Duijnhoven, and F. A. van Dorsten. Assesment of plsda cross validation. *Metabolomics*, 4:81–89, 2008.

[227] D. White. *The Physiology and Biochemistry of Prokaryotes. Second edition.* Oxford University Press. Oxford - New York., 2000.

[228] W. Wiechert. 13c metabolic flux analysis. *Metab Eng*, 3(3):195–206, Jul 2001.

[229] B. M. Wise, N. B. Gallagher, R. Bro, J. M. Shaver, W. Windig, and R. S. Koch. Pls toolbox 5.2. for use with matlab®. Eigenvector Research Inc., Manson., 2005.

[230] J. Wu, N. S. Zhang, A. Hayes, K. Panoutsopoulou, and S. G. Oliver. Global analysis of nutrient control of gene expression in saccharomyces cerevisiae during growth and starvation. *Proceedings of the National Academy of Sciences of the United States of America*, 101(9):3148–3153, 2004.

[231] L. Wu, M. R. Mashego, J. C. van Dam, A. M. Proell, J. L. Vinke, C. Ras, W. A. van Winden, W. M. van Gulik, and J. J. Heijnen. Quantitative analysis of the microbial metabolome by isotope dilution mass spectrometry using uniformly c-13-labeled cell extracts as internal standards. *Analytical Biochemistry*, 336(2):164–171, 2005.

[232] J. G. Xia and D. S. Wishart. Msea: a web-based tool to identify biologically meaningful patterns in quantitative metabolomic data. *Nucleic Acids Research*, 38:W71–W77, 2010.

[233] H. Yamamoto, H. Yamaji, E. Fukusaki, H. Ohno, and H. Fukuda. Canonical correlation analysis for multivariate regression and its application to metabolic fingerprinting. *Biochemical Engineering Journal*, 40(2):199 – 204, 2008.

[234] K. Yizhak, T. Benyamini, W. Liebermeister, E. Ruppin, and T. Shlomi. Integrating quantitative proteomics and metabolomics with a genome-scale metabolic network model. *Bioinformatics*, 26(12):i255–i260, Jun 2010.

[235] N. Zamboni and U. Sauer. Novel biological insights through metabolomics and 13c-flux analysis. *Curr Opin Microbiol*, 12(5):553–558, Oct 2009.

[236] G.-F. Zhang, S. Sadhukhan, G. P. Tochtrop, and H. Brunengraber. Metabolomics, pathway regulation, and pathway discovery. *J Biol Chem*, 286(27):23631–23635, Jul 2011.

[237] J. Zhao and K. Shimizu. Metabolic flux analysis of escherichia coli k12 grown on c-13-labeled acetate and glucose using gg-ms and powerful flux calculation method. *Journal of Biotechnology*, 101(2):101–117, 2003.

[238] Q. Zhou, D. Wang, and M. Xiong. Dynamic flux balance analysis of metabolic networks using the penalty function methods. *SMC IEEE*, pages 3594–3599, 2007.