Network inference from time-resolved metabolomics data

Hendrickx, D.M.

Publication date
2013

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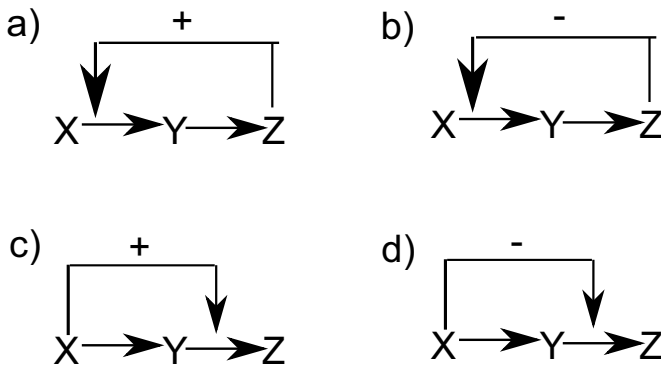
Figure 1.1: Metabolic control mechanisms a) positive feedback control b) negative feedback control c) positive feedforward control d) negative feed forward control

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].
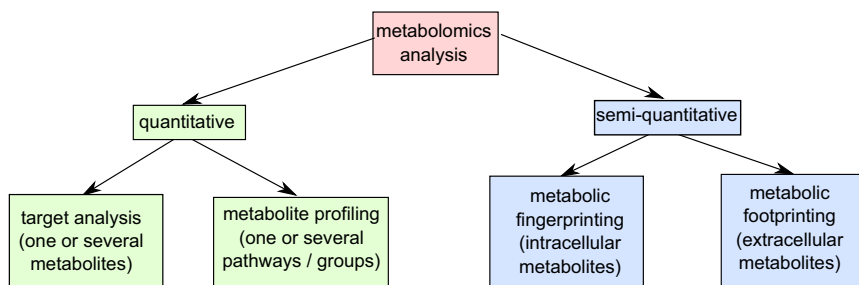

Figure 1.2: Types of measurements in metabolomics.

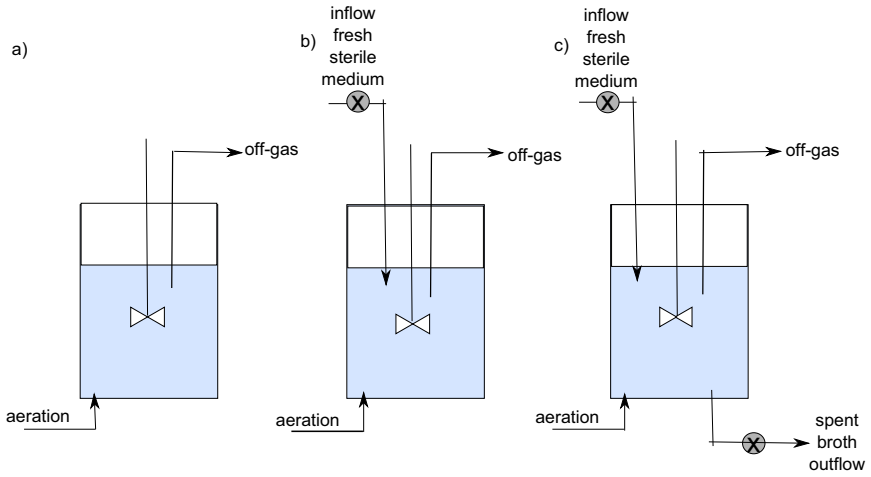## 1.1.2 Sampling methods for time-resolved microbial metabolomics

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)).
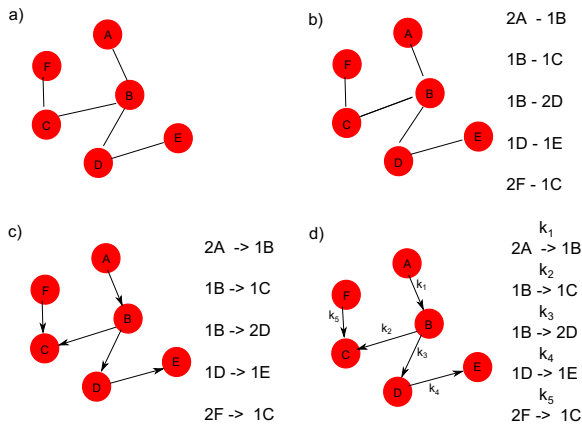

Figure 1.4: Different levels of studying metabolic pathways. a) topology b) stoichiometry c) directionality d) kinetics

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.
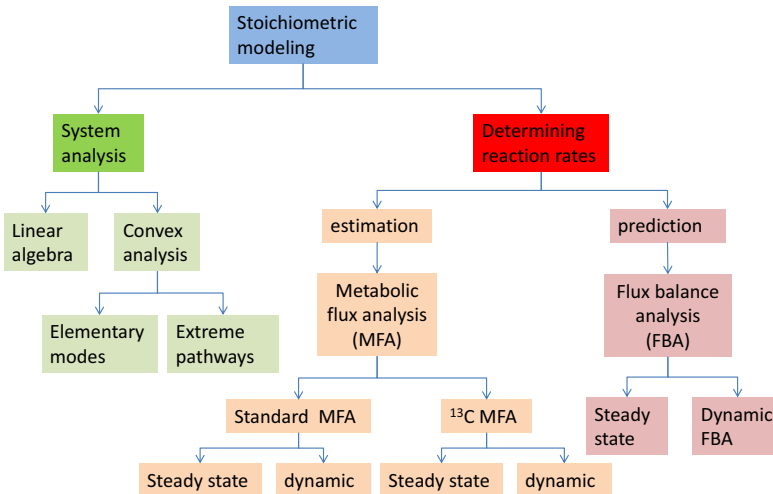

Figure 1.5: Current methodologies in stoichiometric modeling.

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