Metabolic network modeling: Rate Laws and Parameter Estimation

Mathijs Sanders
1340085

Supervisors

Dr. Ir. Dick de Ridder

Information and Communication Theory Group
Bioinformatics
TU Delft
**Introduction**

Metabolism is a vital cellular process and its malfunction is a major contributor to human disease. For this reason, it is important to construct and investigate metabolic networks. We can safely say that such metabolic networks are complex and highly interconnected, therefore system-level computational approaches are required to understand metabolic genotype-phenotype relationships. One of the major fields of research in systems biology is the reconstruction of these metabolic networks [1-2]. The reconstructed network can be used to suggest potential alternatives to known drug targets or could reveal the effects and causes of diseases and therapies.

To understand the complex behavior of the system we need to translate the metabolic network into a dynamical model with rate laws for each enzymatic reaction. These rate laws are defined as mathematical expressions which heavily depend on the underlying mechanism of the enzymatic reactions and can become quite complex with a large quantity of parameters. There are two main reasons to use dynamical models: (1) we wish to gain more knowledge about the system (2) we want to control the system. To model the system as accurately as possible, we wish to have a complete and accurate set of parameters which characterize the system. This definition is by the fact that some parameters are: (1) corrupted by measurement noise (2) influenced by biological variability (3) completely unknown. All these factors of uncertainty and missing knowledge lead to problems in the dynamic modeling of metabolic networks.

To account for these issues we could perform model parameter estimation, also called system identification. With the use of measurement data and the model structure we can estimate the parameters as accurately as possible, which eventually leads to a dynamic model that can make accurate predictions and would be a gain for fundamental research in various fields.

This document gives an overview of the state of the art in the field of system identification in metabolic engineering. First the reader will be introduced to the concept of metabolic engineering after which the reader will be further instructed about its goals, procedures and limitations.

Next the reader will be introduced to the field of metabolic networks and dynamic modeling. The reader will be instructed about its goals, procedures and theoretical methods to extract quantitative and qualitative information.

To translate the metabolic network into a dynamic model one has to define rate laws for each reaction and is explained in Chapter 4. The most well-known and topically important rate laws will be explained in full detail. Not only shall we derive the mathematical expressions for these rate laws, but we will also spend effort on the epiphany of the method in terms of assumptions and underlying biological ideas. The reader will not only see the mathematical derivation, but can also see the types of reaction mechanisms we are trying to model.

Although much information of the metabolic model and its dynamics has been gathered over time, they are still handicapped by uncertain model parameters. As already explained, many limitations can lead to inaccurate model predications with respect to the observed measurements. To account for these issues we introduce in Chapter 5 different methods to estimate model parameters.

Chapter 6 discusses the limitations of the described methods, together with the open problems and directions of future work. Next we propose to combine some complementing methods to form a synergy. Finally a discussion and conclusion will be given.

M.A. Sanders
Metabolic Engineering

2.1 Introduction

The fundamental goal of the methods discussed later is to accurately model metabolism such that we can understand its behavior and improve it by engineering. Here we define metabolic engineering, its goals and applications. There are external influences on metabolism that are not directly obvious and shall be addressed in this chapter. Furthermore there are currently some limitations and challenges in metabolic engineering that we are trying to overcome or resolve. At the heart of metabolic engineering lies the measurement of parameters to reconstruct the dynamic behavior of the metabolic model.

2.2 Metabolism

Metabolism is defined as the total of all chemical reactions that are carried out in an organism. These chemical reactions are catalyzed by enzymes and change the structure of one or more chemical compounds, also called metabolites. These changes are also called biotransformations. A sequence of biotransformations is called a metabolic pathway [3] as illustrated in Figure 2.1.

![Figure 2.1: A metabolic pathway where multiple biotransformations take place. The substrate of one reaction is the product of a previous reaction.](image)

Input metabolites of a metabolic pathway are called substrates and output metabolites are called products. Metabolites can originate from ingestion of food, but can also be products of other metabolic pathways. It already seems obvious that enzymes, as biological catalysts, have a major influence on the metabolism.

For a metabolic pathway to operate efficiently, its activity must be coordinated and regulated by the cell. It seems very unnecessary to synthesize a compound which is already plenty present. Regulation of metabolic pathways depends on elegant mechanisms. Figure 2.2(a) shows a metabolic pathway catalyzed by three enzymes. This metabolic pathway has no feedback mechanism, but as shown in Figure 2.2(b) the end product of this pathway influences the activity of the first enzyme. It binds to an allosteric site on the enzyme that catalyzes the first reaction, resulting in the regulation of the enzyme
activity. Shutting down the first reaction in the pathway efficiently shuts down the whole pathway. This mode of regulation is also called feedback inhibition:

Figure 2.2: (a) A biochemical pathway without feedback inhibition. (b) A biochemical pathway in which the end product is an allosteric inhibitor of the first enzyme in the pathway.

2.3 Metabolic Engineering

The phenotype of the cell, i.e. its appearance and functioning, is to a large extent determined by its metabolism. Manipulating the regulation of metabolic pathways through different components of metabolism, such as enzyme and substrate concentrations, gives bioengineers tools to improve the cellular properties. Although metabolic engineering is a young field, many laboratories have succeeded in reconstructing metabolic pathways in silico. Individual enzymes, metabolites, their respective interactions and reaction mechanisms involved in these pathways have been studied for many years and results are stored in different databases, e.g.:

- **BRENDA** (http://www.brenda-enzymes.info/)

To improve cellular properties or research cellular behavior, researchers make use of genetic engineering (gene duplications or deletions) to alter the gene expression levels which influence the metabolism. Manipulation of one specific enzyme is also called gene targeting. A major problem is that most enzymes participate in multiple pathways, thus a manipulation of an enzyme’s concentration level will have effect on multiple pathways, which could result in undesired side effects. To account for these issues, researchers have developed a mathematical tool called Metabolic Control Analysis, which puts emphasis on the distribution of control over the enzymes in specific pathways. This allows us to infer whether targeting a particular gene will influence other pathways. In the supplement, section S3.3, a full description of Metabolic Control Analysis is given.

These interventions and methods allow us to better understand the genotype-phenotype relationship. In most cases the engineers make use of a strain, such as *Escherichia coli*, to modify. Analysis of the recombinant strain should be applied to see if the performance improved with respect to the original strain. Often, such an analysis will indicate an additional genetic modification is required to further
improve performance. Researchers apply analysis and genetic engineering recursively until a desired performance level is achieved:

- **Analysis:** Analyze a pathway and gain knowledge on how it influences the overall cell function and the desired property.
- **Design:** Find the next target for genetic engineering.
- **Synthesis:** Perform genetic modifications to construct the recombinant strain with improved properties.

### 2.4 Goals

The goals of metabolic engineering are to improve certain cellular activities or to study the behavior of the cell after modifications. These goals are rather abstract; below we give some elaborated goals [1-4].

- **Study of behavior:** By studying the behavior of the metabolic pathways after genetic modifications one could suggest alternatives to known drug targets or reveal the effects and causes of diseases and therapies.
- **Heterologous protein production:** A specific genetic sequence from one organism is inserted into the DNA of another, hence the name heterologous, to create a specific protein. Examples are the production of pharmaceutical proteins (hormones, antibodies, etc.) and novel enzymes. An example is the product of human insulin by a recombinant *E.Coli* strain.
- **Improvement of productivity:** In many industrial processes it is important to continuously improve productivity. This can be achieved by increasing the biosynthetic pathway activity, e.g. by inserting additional gene copies. This will not always work as some metabolic pathways involve many enzymes and the increase of activity of one enzyme does not necessarily result in a higher productivity. In these cases Metabolic Control Analysis can help.
- **New product construction:** It could be interesting to use another host to produce different products. This can be achieved by extending existing metabolic pathways by recruiting heterologous enzymes.
- **Reduction of by-product formation:** In many industrial processes by-products are formed. If these are toxic, they may interfere other metabolic pathways or with the purification of the product. Metabolic engineering may reduce the concentration of the by-product.
- **Substrate utility:** It could be interesting to extend the number of substrates which the organism can utilize in order to more efficiently utilize raw material. Inserting a pathway processing the substrate of interest could however lead to (potentially fatal) by-product formation.
2.5 The influence of other information on metabolic engineering

Various types of information have had an influence on metabolic engineering and changed the views in this field of research [1],[5]:

- **Genome sequences:** Over the last decades, the full genomes of multiple organisms have been sequenced. The availability of the full genome allows metabolic engineers to identify genes participating in metabolic pathways together with their regulatory elements.

- **Gene regulation:** Methods to measure and infer networks, such as the yeast two-hybrid system, allow researchers to predict what the impact of a genetic modification on other genes and on the metabolic pathway will be.

- **Control driven modulation:** Control of the reactions rates along a pathway may be distributed along the enzymes, as elucidated by Metabolic Control Analysis. This means that when one tries to optimize the performance of a metabolic pathway, one has to modify the gene expression levels of multiple enzymes.

- **Metabolic network modeling:** Many tools have been developed to analyze metabolic networks. Although not all kinetic information is currently available, the structure based on its stoichiometry can already provide a direction for modulating metabolism. This type of “non-dynamic” model neglects regulation and control information, but it can already be used to make predictions by methods such as Flux Balance Analysis (FBA). Figure 2.3 shows the typical construction of such an *in silico* model.

- **Gene expression microarrays:** Measuring gene expression with microarrays can give an indication what the enzyme levels are at certain time points. A challenge is that the mRNA levels measured are not the same as the enzyme levels, due to post-transcriptional and post-translational processes.

- **Metabolic concentration levels:** Measuring the concentration levels of metabolites can be done by Mass Spectrometry, section S2.1 in the supplement has been dedicated to this procedure. We ultimately want to know if the alterations have affect on the metabolite concentrations.

![Figure 2.3: In silico model construction](image-url)
2.6 Limitations and challenges

Recent progress in the field of metabolic engineering has been tremendous. Many different recombinant strains have been designed to produce substances needed by industry: pharmaceuticals, fuels, food ingredients, etc. We now give some limitations on metabolic engineering:

- **Complexity and connectivity:** A major problem in metabolic engineering is that the metabolism is highly complex and interconnected. We can measure many properties of the metabolic system, but we generally do not know what the full mechanism behind its regulation is.
- **Regulation and control:** Although there are metabolic models based on bibliomic data, estimations and simulations, these still leave out the regulation and control. A challenge for the future is to combine metabolic models with gene regulation models to gain insight into the mechanism behind pathways, thriving on new and improved measurement techniques.
- **Missing genomic information:** A significant challenge is that we may have fully sequenced the genome, but still not have assigned all ORFs a function. Even for well-modeled organisms such as *Saccharomyces cerevisiae* and *Escherichia coli* only 50% of the ORF’s have an assigned function [1].
3.1 Introduction

With the availability of the genome sequence and its annotation we can attempt to define all metabolic enzymes. In addition many metabolic genes and enzymes have been studied, resulting in a collective knowledge base, including reactions mechanisms and characterized interactions. Manual component-by-component reconstruction of genomic and enzymatic data will lead to a structured reconstruction, also called a metabolic network. This model can be used to compute allowable network states under governing chemical and genetic constraints.

3.2 Reconstructing the metabolic network

The first step is to identify genes from the annotation of the genomic sequence, followed by their functional assignment using various experimental methods and bioinformatics tools (e.g. BLAST). Once we have derived the biological components from the genome sequence, we would like to infer the interactions between them. Next we should define the metabolic specificity for all enzymes. Although high gene- and protein-sequence homology implies a similar function for gene products, this should be experimentally verified. Enzymes can be classified into two groups on the basis of substrate specificity.

1. **Structure specificity**: Those that function on one or a few highly similar substrates.
2. **Broad specificity**: Those that can function on a class of compounds with similar functional groups.

Substrate specificity of enzymes can differ between organisms; this is especially the case for coenzymes. In addition to primary substrate specificity, enzymes from different organisms preferentially use different coenzymes. Since binding sites for coenzymes often share common sequence and structural motifs, consensus sequence motifs can be used to identify coenzyme binding sites based of persevered protein folds [7]. However, primary literature is the most reliable source of information to determine the coenzyme specificity. Next, we need to know the structural formulas of the metabolite, as we need to balance the reaction mechanism to satisfy mass and energy conservation laws. This defines the stoichiometry of the reaction.

Then, the directionality or reversibility of a reaction needs to be specified. Directionality is a function of the thermodynamics of a reaction. To understand the thermodynamics of a biochemical system, section S3.1 about thermodynamics is added to the supplement. Not all biochemical reactions are reversible; modeling this incorrectly can lead to incorrect predictions and behavior. For most reactions, reversibility is specified, but for other reactions calculated thermodynamic properties can be used to determine it. Reactions with a highly negative Gibbs free energy, \( \Delta G \), can be assumed to be irreversible, while those with a Gibbs free energy close to zero can be seen as reversible. The Gibbs free energy of a reaction can be estimated from the structures of the metabolites using a group contribution method [8]. Care needs to be taken as directionality can differ between *in vitro* and *in vivo* due to temperature, pH and metabolite concentration differences. Figure 3.1 shows the steps involved in defining all metabolic reactions.
After defining the metabolic reactions we can assemble the metabolic network by:

- **Analysing traditional biochemical pathways**: These pathways are the fueling reactions of the organism under study. If they cannot construct the building blocks of life (e.g. glucose), some reactions are missing.

- **Filling in missing metabolic activities**: Sometimes metabolic reactions can be found in the organism under study, but the genome sequence does not support them. This is most likely caused by an incomplete annotation of the genome.

Generally, one begins with the assembly of the central metabolism (present in all organisms), and then moves on to the biosynthesis of individual macromolecular building blocks (e.g. amino acids or lipids). Once all the main metabolic pathways are included, several reactions not included in the traditional biochemical pathways, need to be included. The assembled metabolic network can be mathematically represented by the stoichiometric matrix, $N$. This $i \times l$ matrix is defined for the $i$ metabolites and the $l$ reactions. Each column of the matrix represents a reaction whereas its elements are the stoichiometric coefficients:

$$\{N_{il}\} = \begin{cases} 
\text{Stoichiometric coefficient} & \text{If it participates in the reaction} \\
0 & \text{If it doesn't participate in the reaction} 
\end{cases}$$

If the metabolite is a substrate of the reaction its stoichiometric coefficient is negative, if it is a product it is positive. Figure 3.2 shows a set of reactions mathematically represented by the stoichiometric matrix.
3.3 Dynamic models

3.3.1 Introduction

Perturbations of a biochemical system, e.g. by gene modulation or drug treatment, can lead to global effects that are by no means self-evident. To understand and predict its behavior we can make use of systems biology. Systems biology is a biology-based inter-disciplinary study field that focuses on the systematic study of complex interactions in biological systems, using a new perspective (integration instead of reduction) to study them. We do this by computer simulations, for which a mathematical model of the biochemical networks is required. Using only the structure of a metabolic network, we can calculate metabolic fluxes (reaction rates in steady state) by pathway- or constrained-based methods, such as Flux Balance Analysis (FBA). However, such methods do not explain how rates are actually influenced by the activities of enzymes and how they respond to perturbations. Also the stability of the system in chemical equilibrium is not described by this model.

These questions are answered by dynamic or kinetic models, which employ ordinary differential equations (ODE) to describe the temporal behavior of the system in a deterministic way. These models give the researchers a tool to understand and estimate the behavior of dynamically closed or open systems such as metabolism. To illustrate the ODE approach, an example of the GLUT transporter is shown in Figure 3.3.

(A)

![Glucose transport diagram](image)

(B)

![Kinetic diagram](image)

Figure 3.3: (A) The four states of the GLUT transporter. (B) Four-state kinetic diagram of a GLUT transporter.

The rates of the elementary process illustrated in the kinetic diagram, 3.3(B), are determined by the law of mass action. This law states that the rate of a reaction is proportional to the product of the concentrations of the molecular species involved in the reaction. We define $k^+$ as the proportionality constant (rate constant) and the forward rate from one state to the other is given by $J_f = k^+ \prod [G_i]$. It is now easy to write down the system of ODEs. To do so one must keep track of the change that each elementary process makes for each state. Thus the rate of transition of State_1 to State_2 is given by $k_{12}[G]_{out}x_1$, where $x_1$ is the concentration of the first state.
Writing out the system of ODEs gives:

\[
\begin{align*}
\frac{dx_1}{dt} &= -k_{12}[G]_{out}x_1 + k_{21}x_2 - k_{14}x_1 + k_{41}x_4 \\
\frac{dx_2}{dt} &= -k_{21}x_2 + k_{12}[G]_{out}x_1 - k_{23}x_2 + k_{32}x_3 \\
\frac{dx_3}{dt} &= -k_{34}x_3 + k_{43}[G]_{in} - k_{32}x_3 + k_{23}x_2 \\
\frac{dx_4}{dt} &= -k_{41}x_4 + k_{14}x_1 - k_{43}[G]_{in}x_4 + k_{34}x_3
\end{align*}
\]

Using the package XPPAUT (http://www.math.pitt.edu/~bard/xpp/xpp.html) to solve the system of ODEs we constructed Figure 3.4, which shows the temporal behavior of the concentrations. The highest curve to the lowest curve is the concentration of $x_1, x_2, x_3$ and $x_4$ respectively.

![Figure 3.4: Concentration of the molecular species over time.](image)

Concentrations stop changing after a short time period. This is called chemical equilibrium or steady state and indicates that the reaction rates for the production and degradation of the metabolites are equal. The system of ODEs can also conveniently be expressed as:

\[
\frac{d\hat{x}}{dt} = N\tilde{\nu}(\hat{x}, \vec{k})
\]

Where $\hat{x}$ is the concentration vector, $N$ the stoichiometric matrix, $\tilde{\nu}$ the reaction rate vector (in this case expressed by law of mass action) and $\vec{k}$ the parameter vector. Steady state would indicate:

\[
\frac{d\hat{x}}{dt} = N\tilde{\nu}(\hat{x}, \vec{k}) = N\vec{f} = \vec{0}
\]

The reaction rate vector $\tilde{\nu}$ stays constant and in this case also called the flux vector $\vec{f}$. Due to a large variance of enzymatic reactions the reaction rates can also be modeled by different rate laws than the law of mass action. This shall be the topic of Chapter 4.
3.3.2 Thermodynamics and chemical equilibrium

We can describe a chemical reaction as a function of the thermodynamic driving force, ultimately resulting in chemical equilibrium [10-11]. The occurrence of a reaction is also named a natural process or spontaneous change. *Spontaneous change* is that which, once initiated, proceeds on its own until some *state of equilibrium* is attained. One could think of a ball being rolled up a mountain, containing potential energy, and once released this potential energy is converted to kinetic energy ultimately let the ball roll until it halts (the equilibrium state). To further understand this we first introduce some important properties of chemical systems.

Thermal energy or kinetic energy has the tendency to disperse as widely as possible and this is what drives all spontaneous processes. To understand how the direction and extent of the spreading and sharing of thermal energy is related to measurable properties of substances we introduce the entropy:

**Entropy** is a measure of the degree of spreading and sharing of thermal energy within a system. This is also called the disorder of the system.

As a substance becomes more dispersed in space, the thermal energy is also spread over a larger volume, leading to an increase in entropy. Although this holds for ideal gases and cannot be used for all solids or liquids, it turns out that in a dilute solution, the solute can often be treated as a gas dispersed in the volume of the solution. This led to the following definition of the change of entropy as reflected by a change of concentration of $C_1$ to $C_2$.

$$\Delta S = R \ln \left( \frac{C_1}{C_2} \right) \quad (3.3)$$

We define the change of entropy in terms of another quantity namely heat, $q$, under the assumption that the temperature stays constant. For a process that exchanges a quantity of heat $q_{sur}$ with the surroundings, the entropy change is defined as.

$$\Delta S = \frac{q_{sur}}{T} \quad (3.4)$$

Because all natural processes lead to the spreading and sharing of thermal energy and because entropy is a measure of the extent to which energy is dispersed in the world, it follows that

**In any spontaneous macroscopic change, such as chemical reactions, the entropy of the world increases**

This is known as the second law of thermodynamics. The most important entropy in the definition of thermodynamics is not the entropy of the system or surroundings, but the total entropy, called the entropy of the world.

$$\Delta S_{tot} = \Delta S_{sys} + \Delta S_{sur} \quad (3.5)$$

The only way the entropy of the surrounding changes is through exchange of heat with the system:

$$\Delta S_{sur} = \frac{q_{sur}}{T} \quad (3.6)$$
The most fundamental property of a chemical system is its Gibbs free energy. It takes energy to break the chemical bonds that hold the atoms in a molecule together. Thermal energy, because it increases atomic motion (increase in disorder), makes it easier for the atoms to be pulled apart. Both chemical bonding and heat have a significant influence on a molecule; the former reducing disorder, the latter increasing it. The net effect, the amount of energy actually available to break and subsequently form other chemical bonds, is called the free energy of that molecule. More generally, free energy is defined as the energy available to do work in any system. For a molecule within a cell, where pressure and volume usually do not change, the free energy is denoted as $G$. The free energy is defined by the energy contained in the chemical bonds, enthalpy $H$, and the disorder and temperature in degrees Kelvin $T$:

$$G = H - TS \quad (3.7)$$

Chemical reactions break some bonds in the reactants and form new bonds in the products. Consequently, reactions can produce changes in free energy $\Delta G$:

$$\Delta G = \Delta H - T \Delta S \quad (3.8)$$

This change in free energy is a fundamental property of chemical reactions. In some reactions, the change of free energy is positive, indicating that the products of a reaction contain more free energy than its reactants. These so called endergonic reactions do not occur spontaneously as they require an input of energy. Any chemical reaction tends to occur spontaneously if the difference in disorder between the reactants and products is greater than the difference in bond energies. Using (3.5-3.6) we can derive (3.9). To perform the chemical reaction the heat is drawn from the surrounding:

$$\Delta S_{\text{tot}} = \Delta S_{\text{sys}} + \Delta S_{\text{sur}} = -\frac{q_{\text{sur}}}{T} + \Delta S \quad (3.9)$$

This can be written as the change of enthalpy:

$$\Delta S_{\text{tot}} = -\frac{\Delta H}{T} + \Delta S \quad (3.10)$$

Multiplying both sides by $-T$ results in:

$$-T \Delta S_{\text{tot}} = \Delta H - T \Delta S \quad (3.11)$$

and substituting $-T \Delta S_{\text{tot}} = \Delta G$ gives (3.8).

In order to make use of free energies to predict the direction of the spontaneous reactions, we need to know the free energies of the individual components of a reaction. For this purpose we can combine the standard enthalpy of formation and the standard entropy of a substance to get its standard free energy of formation:

$$\Delta G_f^\circ = \Delta H_f^\circ - T \Delta S_f^\circ \quad (3.12)$$

and then determine the standard Gibbs free energy of the reaction according to

$$\Delta G^\circ = \sum_{p \in \text{products}} \Delta G_f^\circ(p) - \sum_{r \in \text{reactants}} \Delta G_f^\circ(r) \quad (3.13)$$
The symbol ° indicates that a quantity is measured under the standard conditions of 1 mole substance and (usually) a temperature of 298 K. For a reaction \( A \leftrightarrow B \), one of the following three situations will always apply:

<table>
<thead>
<tr>
<th>( \Delta G^\circ )</th>
<th>Reaction Proceeds</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt; 0</td>
<td>Forward reaction can spontaneously proceed: ( A \rightarrow B )</td>
</tr>
<tr>
<td>&gt; 0</td>
<td>Reverse reaction can spontaneously proceed: ( A \leftarrow B )</td>
</tr>
<tr>
<td>= 0</td>
<td>the reaction is at <em>equilibrium</em>; the quantities of A and B will not change</td>
</tr>
</tbody>
</table>

The free energy of a substance depends on its concentration. A higher concentration leads to more free energy, thus the conversion of a molecule is accompanied by a fall in free energy. We express the change in free energy when a substance undergoes a change of concentration of \( C_1 \) to \( C_2 \) as:

\[
\Delta G = \Delta H - RT\ln\left(\frac{C_1}{C_2}\right) = 0 - RT\ln\left(\frac{C_1}{C_2}\right) = RT\ln\left(\frac{C_2}{C_1}\right) \quad (3.14)
\]

How can we evaluate the free energy of a specific sample at some arbitrary concentration? First, recall that the standard molar free energy \( G^\circ \) is calculated with respect to a concentration of 1 mole. The free energy per mole of our sample is just the sum of this value and any change in free energy that would occur if the concentration were changed from 1 mole to the concentration of interest, \( C \):

\[
G = G^\circ + RT\ln(C) \quad (3.15)
\]

Under conditions of constant temperature and pressure, chemical change will tend to occur in whatever direction leads to a decrease in the value of the *Gibbs free energy*. When \( G \) falls as far as it can, all net change comes to a stop. Let us assume that we have a reaction where substrates A and B are converted to the products C and D, the free energy of substance as:

\[
G_A = G_A^\circ + RT\ln(C_A) \quad (3.16)
\]

The free energy change for the reaction is given by:

\[
\Delta G = G_C + G_D - G_A - G_B \quad (3.17)
\]

Using equation (3.14) we expand (3.15):

\[
\Delta G = \left( G_C^\circ + RT\ln(C_C) \right) + \left( G_D^\circ + RT\ln(C_D) \right) - \left( G_A^\circ + RT\ln(C_A) \right) - \left( G_B^\circ + RT\ln(C_B) \right) = \Delta G^\circ + RT\ln\left(\frac{C_C C_D}{C_A C_B}\right) = \Delta G^\circ + RT\ln(Q) \quad (3.18)
\]

As a chemical reaction takes place, \( G \) *will only decrease*. Eventually a point is reached where any further transformation of reactants into products would cause \( G \) to increase. As the reaction approaches equilibrium, \( \Delta G \) becomes finally reaches zero. At equilibrium we find \( Q=\text{constant} \), also called the equilibrium constant:

\[
\Delta G^\circ = -RT\ln(K_{eq}) \quad (3.19)
\]
This equation is one of the most important in chemistry as it relates the equilibrium constant to measurable properties of the reactants and products. It is often useful to solve (3.19) for $K_{eq}$:

$$K_{eq} = e^{-\frac{\Delta G^*}{RT}}$$ (3.20)

In kinetics the equilibrium constant is also the ratio of the rate constants of the reactants, for example:

$$K_{eq} = \frac{k_1^+}{k_1^-}$$ (3.21)

A large number of parameter estimation methods try estimating these rate constants. A drawback is that the ratio of the rate constants can change during the estimation, as the relation between the rate constants and the equilibrium constant has not been taken into account, but the equilibrium constant remains constant as the enthalpies and entropies do not change. This implies that one would get an estimated equilibrium constant that is not equal the true equilibrium constant due to thermodynamics, therefore violating the laws of thermodynamics, resulting in a dynamic model that does not occur in reality. These relations pose thermodynamic constraints on the model parameters.

There is an additional thermodynamic constraint called the Wegscheider condition. Section 3.1 was added to the supplement to fully discuss this phenomenon.
3.4 Goals

The goals of the metabolic network and its dynamical model, taken from [1-2] and [16-17], are:

- **Interpretation and evaluation of measured data:** It would be interesting to build a dynamic model that is able to reproduce experimental data and study the behavior of the modeled system.
- **Analysis of properties:** Analysis of the model can help to better understand the behavior of the system. A downside of these methods is that they need a very accurate model, where in practice most parameters of the model are unknown or corrupted by measurement noise. Also stability analysis can be used to analyze properties. Section 3.2 was added to the supplement to fully explain this type of analysis.
- **Simulation:** The simulation of the dynamic model could give insight into the behavior of the system and give explanatory information about its dynamics.
- **Optimisation:** When a metabolic network is defined one can use it to improve the design of a strain such as to have better performances.
- **Prediction:** A major goal of the metabolic networks is to build a validated metabolic network that could predict the outcome of experimentations, such that most calculation can be done *in silico*.

These goals can be divided into two complementing classes:

**Gain knowledge system:** We would like to gain knowledge about the system. With informative analysis we can understand and predict the complex behavior of the system. Some examples of knowledge we could gain are: parameter sensitivity or flux direction.

**Control the system:** We would like to reconstruct the system as accurately as possible to control system. In this case we do not desire to gain any knowledge of the system, but would like that the model reproduces the measurements as accurately as possible.

3.5 Limitations

Although the modeling of metabolic networks has been frequently applied, it still is limited:

- **Unknown model parameters:** Much time has been spent on the determination of model parameters in a metabolic network. Most known parameters are stored in databases, but not all parameters are known. This is a limiting factor as we try to model the *in vivo* system as complete as possible.
- **Uncertain model parameters:** A major obstacle is that model parameters are corrupted by noise.
- **Different conditions:** Some parameters are estimated in experimental configurations different from the one that you are trying to model. Due to biological variability this can lead to inaccurate modeling and leads to inaccurate information and simulations.
- **Missing reactions:** Most of the reactions that play part in our metabolism have been studied and its results stored in databases, but not for all. To account for these issues one could make use of metabolic network alignment in different species to find missing reactions [18].
- **Complexity:** Metabolism is highly complex and hard to model, due to different levels of regulation and high connectivity. Integrating the different levels of regulation has to this day not been fully accomplished, resulting in a simplified model.
Rate Laws

4.1 Introduction

Once the structure of a metabolic network is known it can be translated into a dynamic model, requiring rate laws for all enzymatic reactions. The rate law for a chemical reaction is an equation that links the reaction rate to concentrations of reactants, pressure, temperature and the rate constants. Let us assume that we have the following reaction:

\[ nA + mB \rightarrow C \]

We can make use of the law of mass action to specify the rate equation in terms of the reactants.

\[ \frac{d\chi}{dt} = k(T)[A]^n[B]^m \quad (4.1) \]

Where the function \( k(T) \) is the rate constant, which should be better dubbed reaction rate coefficient as it depends on the temperature, as seen in the Arrhenius equation:

\[ k = Ae^{-\frac{E_a}{RT}} \quad (4.2) \]

where \( k \) is the reaction rate coefficient, \( A \) the pre-factor or frequency factor, \( E_a \) the activation energy, \( R \) the universal gas constant and \( T \) the temperature in Kelvin. The activation energy is the energy that must be overcome so that the reaction occurs. With the addition of a catalyst, such as an enzyme, this energy barrier is much lower as shown in Figure 4.1 and results in a higher reaction rate coefficient.

![Energy barrier for the reaction where the substrate S is converted to the product P.](image)

The rate equation of rate laws is dependent on the underlying enzyme mechanism. With different enzyme mechanisms, one should use different rate laws. A problem is that these rate equations can become quite involved and may contain a large number of parameters, for example: reaction rate coefficients and enzyme concentrations. For most enzymes the enzyme parameters are unknown, which will be a topic of discussion in the next chapter.

Enzyme assays are laboratory procedures that measure the rate of enzymatic reactions. Because enzymes are not consumed by the reactions they catalyse, enzyme assays usually follow changes in the concentration of either substrates or products to measure the rate of reaction. The most well-known assay is the initial rate experiments where the initial rate of a reaction is determined. This is done by looking at the concentration of the product over time as shown in Figure 4.2. This curve has approximately a straight
line in the beginning, called the initial rate period, from which we can calculate the slope. This slope is also called the initial reaction rate used frequently in metabolic modeling.

![Initial rate period](image)

**Figure 4.2: Accumulation of the product over time**

The law of mass action is defined for single-step reactions and its initial reaction rate can be approximated by the substrate concentrations and reaction rate coefficient as seen in (4.1). Its assumptions are also a downside for getting the analytic rate equation. Most of the reactions in metabolism are not performed by a single conversion step and the rate equation does not model the saturable enzyme phenomenon.

### 4.2 Irreversible Michaelis-Menten rate law

A well-known rate law postulated in this section is the irreversible Michaelis-Menten rate law. Let us first introduce the enzymatic mechanism for this rate law:

\[
E + S \xrightleftharpoons{k_1}{k_{-1}} ES \xrightleftharpoons{k_2}{k_{-2}} E + P
\]

Where \(E\) is the enzyme, \(S\) the substrate, \(ES\) the enzyme-substrate complex and \(P\) the product. The algebra needed to deal with a four-reaction system is very complex. To avoid this, Michaelis and Menten assumed that the enzyme is irreversible, but this is frequently incorrect. Instead Briggs and Haldane (1926, UK) limited their derivation to dealing with initial rates of reaction, when no product has formed yet, \([P] = 0\). In this case the rate of binding of product, \(k_{-2}\), is zero, and can be safely ignored. The system of ODEs for this reaction scheme is given by:

\[
\begin{align*}
\frac{d[E]}{dt} &= k_1^{-}[ES] - k_1^{+}[E][S] + k_2[ES] \\
\frac{d[S]}{dt} &= -k_1^{+}[E][S] + k_1^{-}[ES] \\
\frac{d[ES]}{dt} &= k_1^{+}[E][S] - k_2^{-}[ES] - k_2[ES] \\
\frac{d[P]}{dt} &= k_2[ES]
\end{align*}
\]
Figure 4.3 illustrates the temporal behavior of the system of ODEs \((S(0)=0.5, E(0)=0.5)\) and shows that the product trajectory behaves as illustrated in Figure 4.2. The basic principle of the Briggs-Haldane theory is based on the observation that the reaction rate rapidly attains a fixed rate \(v_0\), implying that the concentration \([ES]\) is in Quasi Steady-state. We can state that the initial reaction rate \(v_0\) is then entirely described by reaction step 2, since all P formed must have passed just once through step 2. This implies:

\[
v_0 = k_2[ES] \quad (4.3)
\]

We run into a problem, as at the beginning of the experiment we only know the initial substrate concentration \([S]_0\) and the total enzyme concentration \([E]_0\). The concentration of the enzyme-substrate is normally not measurable, thus we need to make assumptions such that we can define the reaction rate equation by these two available concentrations. Another fact that led to the derivation of the Michaelis-Menten rate law is that when one increases the concentration of the substrate that at a certain point the initial reaction rate does not increase. This fact is shown in Figure 4.4, where slope in the initial period of the product concentration curves does not increase with an increase of substrate at a certain point.

Figure 4.4: Product accumulation trajectories over time corresponding to different substrate concentrations. The substrate concentrations used are: 0.5 mM, 1mM, 5 mM, 10 mM, 15 mM and 20 mM.
This natural phenomenon is called enzyme saturation and happens frequently in reality. For a given enzyme concentration and for relatively low substrate concentrations, the reaction rate increases linearly with substrate concentration; the enzyme molecules are largely free to catalyze the reaction, and increasing substrate concentration indicates an increasing rate. However, at relatively high substrate concentrations, the reaction rate asymptotically approaches the theoretical maximum; the enzyme active sites are almost all occupied and the reaction rate is determined by the intrinsic turnover rate of the enzyme.

The concentration of the enzyme-substrate complex changes much slower than that of the product and substrate, hence changes on a different time scale. For the rate equation we use the most comprehensible derivation made by Higgs and Haldane. The complete derivation can be found in the supplement section S4.1. Under the Quasi-Steady state assumption we define:

\[ \frac{d[ES]}{dt} = k_1[E][S] - [ES](k_1^{-} + k_2) = 0 \quad (4.4) \]

Which we rewrite to:

\[ [ES] = \frac{[E][S]}{K_m} \quad (4.5) \]

Where the constant \( K_m \) is the Michaelis-Menten constant, which shall be discussed later in this section. Using conservation laws (supplement) we can rewrite (4.5):

\[ [ES] = [E]_0 \frac{1}{1 + \frac{K_m}{[S]}} \quad (4.6) \]

Substituting this expression in (4.3).

\[ v_0 = k_2[ES] = k_2[E]_0 \frac{1}{1 + \frac{K_m}{[S]}} = k_2[E]_0 \frac{[S]}{[S] + K_m} = V_{\max} \frac{[S]}{[S] + K_m} \quad (4.7) \]

Giving the rate equation for the Michaelis-Menten reaction scheme. First of all the rate equation is dependent on the initial concentration of the substrate and the total enzyme concentration, which are known at the beginning of the experiment. Figure 26 illustrated that if one would increase the substrate concentration it would lead to an asymptotic limited initial reaction rate due to enzyme saturation. This phenomenon is also modeled by this rate equation as shown in Figure 4.5. The initial reaction rate is bounded to the product of the reaction rate coefficient of step two times the total enzyme concentration as the ratio in (4.7) goes to unity when \([S] \gg K_m\). Such high substrate concentrations is rare, a more appropriate measure to characterize an enzyme is the substrate concentration at which the initial reaction rate reaches half of its maximum value \( \left( \frac{V_{\max}}{2} \right) \) (4.8). This is the case were the substrate concentration is equal to the Michaelis-Menten constant defined by (S4.4).

\[ v_0 |_{[S]=K_m} = V_{\max} \frac{[S]}{[S] + K_m}_{[S]=K_m} = \frac{V_{\max}}{2} \quad (4.8) \]
The Michaelis-Menten rate equation has also some limitations. First of it is assumed that the rate of change of the enzyme-substrate complex is equal to zero, but this is not always the case. The quality of the Michaelis-Menten approximation depends on the timescale separation present in the dynamics, which controls the magnitude of the rate of change of $[ES]$. Another limitation is that the Michaelis-Menten kinetics relies on the laws of mass action, which, in turn, is derived from the assumptions of free diffusion and thermodynamically-driven random collision. This is not always the case as the cytoplasm of a cell is more gel-like liquid due to the protein concentrations resulting in a non-free diffusion and non-random collisions.

### 4.3 Reversible Michaelis-Menten rate law

The reversible Michaelis-Menten is another rate law is commonly used in dynamic models and extends the irreversible Michaelis-Menten rate law. Let us first introduce the reaction scheme.

\[
\begin{align*}
E + S &\rightleftharpoons ES \\
&\rightleftharpoons EP \\
&\rightleftharpoons E + P
\end{align*}
\]

From this reaction scheme we can deduce that there is an extra step, namely the conversion of the enzyme-substrate complex to the enzyme-product complex. We can also see that all the reactions are now reversible. From this reaction scheme we can construct the system of ODEs:

\[
\begin{align*}
\frac{d[E]}{dt} &= -k_1^+ [E][S] + k_1^- [ES] + k_3^+ [EP] - k_3^- [E][P] \\
\frac{d[S]}{dt} &= -k_1^+ [E][S] + k_1^- [ES] \\
\frac{d[ES]}{dt} &= k_1^+ [E][S] - k_1^- [ES] + k_2^+ [EP] - k_2^- [ES] \\
\frac{d[P]}{dt} &= k_3^+ [EP] - k_3^- [E][P]
\end{align*}
\]
To derive the rate equation for this rate law we must make a simplifying assumption.

**Equilibrium assumption**: The dissociation of the enzyme-substrate complex and the enzyme-product complex is much faster than the interconversion between the enzyme-substrate complex and the enzyme-product complex.

This implies that $k_1^- \gg k_2$ and $k_3^+ \gg k_3^-$. Therefore the first step and the third step are near equilibrium, implying that the equilibrium constants or the reciprocal of them, named the dissociation constants, are near their steady-state value. The dissociation constants of the first and the last step are:

$$K_s = \frac{k_1^-}{k_1^+} = \frac{[E][S]}{[ES]} \quad (4.9)$$
$$K_p = \frac{k_3^-}{k_3^+} = \frac{[E][P]}{[EP]} \quad (4.10)$$

The full derivation of the rate equation can be found in the supplement section S4.2. Rewriting (4.9), (4.10).

$$[ES] = [E]_0 \frac{[S]}{K_s} \left( 1 + \frac{[S]}{K_s} \frac{[P]}{K_p} \right) \quad (4.11)$$
$$[EP] = [E]_0 \frac{[S]}{K_s} \left( 1 + \frac{[S]}{K_s} \frac{[P]}{K_p} \right) \quad (4.12)$$

The rate equation for the reversible Michaelis-Menten reaction scheme is:

$$v_0([S], [P]) = k_2^+ [ES] - k_2^- [EP] = \frac{v_f \frac{[S]}{K_s} - v_l \frac{[P]}{K_p}}{1 + \frac{[S]}{K_s} + \frac{[P]}{K_p}} \quad (4.13)$$

Equation (4.13) introduces two new constants, namely the forward rate $v_f$ and the backward rate $v_l$. These are defined as the maximum rate that can be acquired by the forward rate and backward rate just as the definition of $V_{\text{max}}$ in the reversible Michaelis-Menten rate law. This can be seen by setting for instance the concentration of the product to zero which will reduce the equation to:

$$v_0([S], 0) = v_f \left( \frac{[S]}{K_s} \right) = v_f \left( \frac{[S]}{K_s + [S]} \right) \quad (4.14)$$
In this context the dissociation constant behaves almost the same as the Michaelis-Menten constant and can be shown by:

\[ K_s = \frac{k_1^-}{k_1^+} \neq K_m = \frac{k_1^- + k_2^+}{k_1^+} \]

But we made the assumption that \( k_1^- \gg k_2^+ \) such that the constant \( k_2^+ \) can be neglected:

\[ K_s = \frac{k_1^-}{k_1^+} \approx K_m \]

For this rate law the same rules apply for the irreversible Michaelis-Menten kinetics, such as enzyme saturation. Although the reversible Michaelis-Menten kinetics is a better approximation of the behavior of enzymatic reaction rates it still has its limitations. The first limitation is that it does not model the cooperative kinetics between enzymes. A second major limitation is that it fails to explain the allosteric regulation of enzymes (e.g. feedback inhibition).

### 4.5 Convenience kinetics

#### 4.5.1 Introduction

Dynamic modeling of metabolic networks requires information about the enzymatic reactions. To capture these enzymatic mechanisms all the reactions in an enzymatic climate are described by different rate laws, such as the laws of mass action and linlog kinetics [28]. One of the fundamental disadvantages of these rate laws is that they fail to describe the enzyme saturation phenomenon and could be better modeled by irreversible and reversible Michaelis-Menten rate laws. Other rate laws have been derived from specific enzymatic mechanisms, but generally have a complicated mathematical expression and have to be established separately for each reaction stoichiometry. One of the major disadvantages of metabolic networks is that they are dependent on a large number of enzymatic parameters (e.g. equilibrium constants, Michaelis-Menten constants and reaction rate coefficients). To model allosteric regulation one has to gather an additional parameter.

**Modifier constant:** In metabolism some reaction rates are modified by inhibitors or activators. These modifiers bind to the allosteric site of the enzyme and influence its activity; its effect is quantified by modifier constants.

Although we have information about a large number of parameters, using them is not so straightforward. The metabolic networks are handicapped by uncertain or unknown parameters (see chapter 3), but are also constrained by thermodynamic constraints (supplement S3.1).

**Thermodynamic constraints:** Parameters are constrained by the thermodynamics of the system. If one would construct a model which violates the laws of thermodynamics it is likely not to model the in vivo system and produce inaccurate results or behaves differently.

If one would develop methods that estimate the enzyme model parameters, like parameter fitting or optimization, it is likely that the estimated parameters violate the thermodynamic constraints (as it is not taken into account).
4.5.2 Convenience kinetics

To take thermodynamic constraints into account Liebermeister, et al. [29] developed a rate law called convenience kinetics, based on thermodynamically independent parameters, which is a direct generalization of the reversible Michaelis-Menten rate law. One of the advantages is that it can describe all possible stoichiometries and enzyme regulation by modifiers. First we should state the assumptions of this rate law:

- **Random order**: The substrates bind to the enzyme in an arbitrary order and are converted in products, which dissociate from the enzyme in an arbitrary order.
- **Rapid-equilibrium**: The binding of substrates and products is reversible and much faster than the conversion step. This is the same as in reversible Michaelis-Menten.
- **Reversibility**: All reactions in the reaction scheme are assumed to be reversible.
- **Binding energy**: The binding energy of individual reactants does not depend on reactants already bound to the enzyme.

For the last assumption we should explain the term “binding energy”. The interaction between the substrate and the enzyme hold the substrate in the proper orientation for the most effective catalysis. The energy derived from this interaction is called the binding energy and used to lower the activation energy barrier, and stabilize the transition state from the enzyme-substrate complex to the enzyme-product complex. This binding energy can be calculated using standard Gibbs free energy. We shall derive the convenience rate equation from a bimolecular reaction without enzyme regulation. Under the assumptions made the reaction scheme would look like Figure 4.6.

**Figure 4.6: Reaction mechanism of a bimolecular reaction under convenience kinetics.**

The dissociation constant

\[ K_A^M = \frac{[A][E_0]}{[E_A]} \quad (4.15) \]

describes the balance of bound and unbound A in chemical equilibrium and is the reciprocal of the equilibrium constant. We have also seen from (3.17-3.18) that the equilibrium constant can be expressed by standard Gibbs free energies automatically implying that the dissociation constant can also be expressed in standard Gibbs free energies (4.16).

\[ K_A^M = \frac{1}{K_{eq}} = \frac{1}{e^{-\Delta G^\circ}} = \frac{1}{e^{-\left(\frac{G^\circ_{E_A} - G^\circ_{A} - G^\circ_E}{RT}\right)}} = e^{-\left(\frac{\Delta G^\circ}{RT}\right)} \quad (4.16) \]
With binding energy assumption of $A$ we can also make analogous assumptions for the binding of $X$. First we shall use the following abbreviations for the sake of simplicity:

$$[\tilde{A}] = \frac{[A]}{K^M_A}, \quad [\tilde{X}] = \frac{[X]}{K^M_X}$$

With the simplifying assumption of independence of binding energy we get:

$$K^M_A = \frac{[E_0][A]}{[E_A]} \Rightarrow [E_A] = \frac{[A]}{K^M_A}[E_0] = [\tilde{A}][E_0] \quad (4.17)$$

$$K^M_X = \frac{[E_0][X]}{[E_X]} \Rightarrow [E_X] = \frac{[X]}{K^M_X}[E_0] = [\tilde{X}][E_0] \quad (4.18)$$

$$K^M_{AX} = \frac{[E_A][X]}{[E_{AX}]} \Rightarrow [E_{AX}] = [\tilde{X}][E_A] = [\tilde{A}][\tilde{X}][E_0] \quad (4.19)$$

This can be done for all enzyme complexes. The total enzyme concentration $E$ is the sum over the concentrations of all enzyme complexes and the free enzyme, in our example:

$$[E] = [E_0](1 + [\tilde{A}] + [\tilde{X}] + [\tilde{B}] + [\tilde{Y}] + [\tilde{A}][\tilde{X}] + [\tilde{B}][\tilde{Y}]) \quad (4.20)$$

We made the assumption that some the interconversion step is much slower than the dissociation reactions. With the laws of mass action and the assumptions we can construct the rate equation, the full derivation can be found in the supplement, section S4.3.1:

$$v_0([A], [X], [B], [Y]) = \frac{k^\text{cat}_{AX}[E_{AX}] - k^\text{cat}_{BY}[E_{BY}]}{1 + [\tilde{A}]} \frac{k^\text{cat}_A[A][\tilde{X}] - k^\text{cat}_B[\tilde{B}][\tilde{Y}]}{1 + [\tilde{X}]} \quad (4.21)$$

In the supplement section S4.3.6 one can find an additional example illustrated for a larger reaction scheme. With the example and assumptions we now formulate the rate equation for all possible stoichiometries and use the vector $\tilde{a}$ and $\tilde{b}$ as the substrate and product concentration vector, respectively.

$$v_0(\tilde{a}, \tilde{b}) = [E] \frac{k^\text{cat}_{AX}\prod_i[\tilde{A}_i]^{a_i} - k^\text{cat}_{BY}\prod_j[\tilde{B}_j]^{\beta_j}}{\prod_i(1 + [\tilde{A}_i]) + \cdots + [\tilde{A}_i]^{a_i} + \prod_j(1 + [\tilde{B}_j]) + \cdots + [\tilde{B}_j]^{\beta_j}) - 1}$$

$$= [E] \frac{k^\text{cat}_A\prod_i[\tilde{A}_i]^{a_i} - k^\text{cat}_B\prod_j[\tilde{B}_j]^{\beta_j}}{\prod_i(\Sigma_{m=0}[\tilde{A}_i]^m) + \prod_j(\Sigma_{m=0}[\tilde{B}_j]^m) - 1} \quad (4.22)$$
We could also model the modifiers in the rate equation, using the assumption that the Gibbs free energy for binding does not depend on the binding of other modifiers and can be used as prefactor for the rate equation. The functions for the modifiers are given by:

$$h_A([D], K^A) = 1 + \frac{[D]}{K^A} \quad \text{Activator function}$$

$$h_I([D], K^I) = \frac{[D]}{K^I + [D]} \quad \text{Inhibitor function}$$

where $[D]$ is the concentration of the modifier, $K^I$ the inhibitor constant and $K^A$ the activator constant. To incorporate the modifiers into the rate equation we can parameterize the metabolic network by the stoichiometric matrix $\mathbf{N}$ and the regulation matrix $\mathbf{W}$ (for the modifiers). If a substance $i$ interacts in a reaction $l$ we can state that its stoichiometric coefficient $n_{il}$ is not zero and comes with its own dissociation constant $K^M_{il}$. If a modifier $i$ interacts in the reaction $l$ as an activator the regulation coefficient is $w_{li} = 1$ if the modifier inhibits it’s indicated by $w_{li} = -1$, otherwise $w_{li} = 0$. We can rearrange the substrate and product vector into one concentration vector $\tilde{c}$ for ease of notation. The rate equation is now defined by:

$$v_l(\tilde{c}, N^+, N^-, W^+, W^-) = E_l \prod_m h_A([C_m], k^A_{lm})^{w^A_{lm}} h_I([C_m], k^I_{lm})^{w^I_{lm}} - \frac{k_{cat}^A \prod_i [\tilde{C}_{il}]^{n^A_{il}} - k_{cat}^I \prod_i [\tilde{C}_{il}]^{n^I_{il}}}{\prod_i \left(\sum_{m=0}^{n_{il}} [\tilde{C}_{im}]^m\right) + \prod_j \left(\sum_{m=0}^{n_{il}} [\tilde{C}_{jm}]^m\right) - 1}$$ (4.23)

with the abbreviation $[\tilde{C}_{il}] = \frac{[C_i]}{K^M_{il}}$. For ease of notation the authors defined the matrices $N^+ = \{n^+_{il}\}$, $N^- = \{n^-_{il}\}$ which respectively contain the absolute values of all positive and negative elements of $\mathbf{N}$, separating the substrates from the products. The matrices $W^+$ and $W^-$ are derived from $\mathbf{W}$ in the same way.

### 4.5.3 Thermodynamic dependence between parameters

The convenience rate law has one major drawback: its parameters are constrained by thermodynamic laws. By setting (4.23) to zero, we obtain the Haldane relationship for the convenience kinetics. This full derivation can be found in the supplement, section S4.3.2. The equilibrium constant for reaction $l$ is defined as:

$$K_{eq} = \prod_j [B_j]^{\beta_j} \prod_i [A_i]^{\alpha_i} = \frac{k_{cat}^A \prod_j (K^M_{bj})^{\beta_j}}{k_{cat}^I \prod_i (K^M_{ai})^{\alpha_i}}$$ (4.24)

Taking the natural logarithm of (4.24) results in:

$$\ln(K_{eq}^l) = \ln(k_{cat}^A) - \ln(k_{cat}^I) + \sum_i n_{il} \ln \left(\frac{K^M_{il}}{K_{il}}\right)$$ (4.25)

Please pay attention to the fact that the last term in (4.29) comes from the properties of the natural logarithm of (4.24). We also know that the equilibrium constant is related to the standard Gibbs free energies of the reactants (4.25) (derivation supplement section S4.3.3):

$$\ln(K_{eq}^l) = -\frac{\sum_i n_{il} G_i^r}{RT}$$ (4.25)
Equations (4.24) and (4.25) imply that the model parameters are constrained by thermodynamic laws through the equilibrium constant, as it relates the reaction rate coefficients to molecular energies. It implies that parameters (through the Wegscheider condition) in the entire network are coupled; an arbitrary choice can easily violate the second law of thermodynamics, which is a severe obstacle to parameters optimisation and fitting. We should devise methods such that we can change the model parameters without violating this constraint.

### 4.5.4 Thermodynamically independent system parameters

To circumvent this problem Liebermeister, et al. [29] introduced new thermodynamically independent system parameters. For a substance $i$ they defined the energy constant:

$$ K_i^G = e^{G_i^* / R T} \quad (4.26) $$

Based upon the Gibbs free energy of formation for substance $i$ we can calculate the energy constant, but can also be estimated from the molecular structure [30]. Another thermodynamically independent system parameter introduced is the velocity constant:

$$ K_i^V = (k_{i+}^{cat} k_{i-}^{cat})^{1 / 2} \quad (4.27) $$

They propose to use these energy and velocity constants as model parameters and treat the equilibrium constants and reaction rate coefficients as dependent variables to circumvent the violation of thermodynamic constraints. The equilibrium constants are calculated from (4.25) and the reaction rate coefficients are defined such that they satisfy (4.24). The calculation of the reaction rate coefficients can be done by the following formula:

$$ k_{i+}^{cat} = k_i^V \prod_k (K_l^G K_{il}^M)^{T_{il}/2} \quad (4.28) $$

By taking the logarithm of both sides we get:

$$ \ln(k_{i+}^{cat}) = \ln(k_i^V) + \frac{1}{2} \sum_l n_{il} \left( \ln(K_l^G) + \ln(K_{il}^M) \right) \quad (4.29) $$

The derivation of these expressions can be found in the supplement, section S4.3.4. We can now define linear equations for all dependent variables from the thermodynamically independent system variables as listed in Table 4.1.
Table 4.1: Linear equations to calculate the defined dependent model parameters

<table>
<thead>
<tr>
<th>Quantity</th>
<th>Symbol</th>
<th>unit</th>
<th>Formula</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gibbs free energy of formation</td>
<td>$G_i^{(0)}$</td>
<td>kJ/mol</td>
<td>$G_i^{(0)} = RT \ln h_i^{(0)}$</td>
</tr>
<tr>
<td>Gibbs' ex. for substrate binding</td>
<td>$\Delta G_i^{(0)}$</td>
<td>kJ/mol</td>
<td>$\Delta G_i^{(0)} = RT \ln h_i^M$</td>
</tr>
<tr>
<td>Equilibrium constant</td>
<td>$k_i^{eq}$</td>
<td>–</td>
<td>$\ln k_i^{eq} = -\sum_i n_i \ln h_i^{(0)}$</td>
</tr>
<tr>
<td>Turnover rate</td>
<td>$k_i^{eq}$</td>
<td>1/s</td>
<td>$\ln k_i^{eq} = \ln h_i^{eq} + \frac{1}{2} \sum_i n_i (\ln h_i^{(0)} + \ln h_i^M)$</td>
</tr>
<tr>
<td>Maximal velocity</td>
<td>$v_{max}$</td>
<td>m mole/s</td>
<td>$\ln v_{max} = -\ln E_i + \ln h_i^{eq} + \frac{1}{2} \sum_i n_i (\ln h_i^{(0)} + \ln h_i^M)$</td>
</tr>
</tbody>
</table>

We can conveniently write the linear equations as a matrix multiplication. Let $\hat{\theta}$ denote the vector of logarithmic independent system parameters and $\hat{x}$ denote the vector containing logarithmic dependent model parameters. We could easily construct the matrix $R_\theta^x$ from the network structure and the equations from Table 1. For the sake brevity of this will not be explained here and one should look at the additional file of [29].

$$\hat{x}(\theta) = R_\theta^x \hat{\theta} \quad (4.30)$$

By inserting (4.28) into (4.22) we obtain a rate law in which all parameters can be varied independently, remaining in accordance with thermodynamics. In its thermodynamically independent form, the convenience kinetics reads

$$v_i = E_i \prod_m h_A([C_m], k_{fm}^{A}) w_{im} h_I([C_m], k_{Im}^{M}) w_{im} K_i^Y \frac{\prod_l [\tilde{C}_{il}]^{n_{il}} (\tilde{K}_{il}^M)^{n_{il}}}{\prod_l \left( \sum_{m=0}^{n_{il}} [\tilde{C}_{il}]^m \right) + \prod_{j} \left( \sum_{m=0}^{n_{ij}} [\tilde{C}_{ij}]^m \right) - 1} \quad (4.31)$$

where $\tilde{K}_{il}^M = K_l^{\tilde{G}} K_i^{M}$.

### 4.5.5 Independent equilibrium constants as system parameters

The energy constants are constructed for two reasons.

1. They provide a consistent way to describe the equilibrium constants
2. If Gibbs free energies of formation are known from experiments or estimations, they can be used for fitting the energy constants and will contribute to a good choice of equilibrium constants

A downside is that not always the Gibbs free energy of formation is known [31], and a different choice of system parameters may be appropriate. If the stoichiometric matrix N has full column rank than the equilibrium constants are independent and (4.25) can always be satisfied by some choice of the $G_i^*$. If this is not the case we should choose a set of reactions with the following property: their equilibrium constants are thermodynamically independent and they determine all other equilibrium constants in the linear model via:

$$\ln(\tilde{k}^{eq}) = R_{ind}^{eq} \ln(\tilde{k}^{ind}) \quad (4.32)$$
How to find this set of independent reactions and how to calculate the equilibrium constants can be found in the supplement, section S4.3.5. Using the newly calculated set of independent equilibrium constants we can find some choice of $G_i^\circ$.

### 4.5.6 Parameter estimation

Given a metabolic network, one could mine the literature for thermodynamic and kinetic data and place their logarithmic values in the vector $x^\ast$. We should now try to find a vector $\hat{\theta}$ of logarithmic independent system parameters that yields a good match between the resulting parameter predictions $\hat{x}(\theta)$ and the data $\tilde{x}^\ast$. We can do this by the method of least squares to yield an estimate of the system parameters. By using (4.30) again we get consistent values of all kinetic parameters from the system parameters. Thermodynamical dependence in the original data is resolved.

### 4.5.7 Conclusion and limitations

Although convenience kinetics has some convenient properties, the assumptions might be too simplistic and could result in a wrong modeling of specific reactions.

**Binding order assumption:** It is assumed that that the substrates bind in arbitrary order, but this is not the case for all enzymatic reactions. In some enzymatic reactions another substrate must bind, changing the conformation of the enzyme opening the binding site for another substrate. In this case the convenience kinetics is a too simplified rate law.

**Binding energy assumption:** Furthermore it is assumed that the binding energy of a reactant does not depend on the binding of another substrate earlier in the process. In some cases the binding of a substrate leads to a change in the conformation in the enzyme which automatically implies that the binding energy for other substrates might be changed.

**Velocity constant:** The velocity constant is calculated from the reaction rate coefficients of a particular reaction (4.32). This is a bit odd as it is the reaction rate coefficients that we are trying to estimate through the independent parameters. One can change thermodynamically independent parameters without violating the constraints, but it does not provide the proof that the calculated dependent parameters are close to the true ones.

**Inheritance:** The convenience kinetics is a generalization of the reversible Michaelis-Menten kinetics resulting in the same limitations.

Convenience kinetics is a mathematical handy and biological plausible choice whenever the detailed enzymatic mechanism is unknown. It can be used for modeling biochemical systems in a simple and standardized ways and also incorporates the natural phenomenon of saturation. It furthermore allows the modeling of regulation by modifiers. This method tackles the problem of thermodynamic constraints which leads to linear dependencies between the logarithmic parameters. To eliminate these constraints a set of thermodynamically independent system parameters is used which can calculate the dependent variables based on the Gibbs free energy of formation readily available in the literature. The basic idea is that parameter estimation methods using convenience kinetics can change the thermodynamically independent parameters such that the dependent parameters do not violate the thermodynamic constraints.
Model parameter estimation

5.1 Introduction

Previous chapters have discussed how metabolic networks can be modeled by a system of ordinary differential equations describing the kinetics of enzymatic reactions. Unfortunately, there is a disproportion between the large number of parameters contained in these rate equations and the relatively small amount of data available. Most of the parameters experimentally determined are uncertain or have been measured under different environmental conditions and are in some cases even unknown.

This is the reason why parameter estimation in biochemical models, such as the metabolic networks, is an integral part of systems biology. To cope with these limitations we have to estimate the parameters on the basis of available data. A problem arising is that most experimental data is sparse, corrupted by measurement noise and dispersed over different databases. Therefore a number of new methods have been developed to estimate the parameters and shall be described in this chapter. It is hard to classify the methods, but a distinction we can make is that some parameters are estimated by a bottom-up type of modeling whereas other methods try to estimate the parameters in a top-down type of modeling.

**Bottom-up modeling:** Traditional modeling methods in biochemistry and cellular biology are bottom-up approaches where models are constructed by piecing together information about individual components (usually obtained in vitro). In a bottom-up approach the individual base elements of the system are first specified in great detail. These elements are then linked together to form larger subsystems, which then in turn are linked, sometimes in many levels, until a complete top-level system is formed.

**Top-down modeling:** The systems biology approach, however, lends itself to top-down (black box) approaches, breaking down a system to gain insight into its compositional sub-systems. In a top-down approach an overview of the system is first formulated, specifying but not detailing any first-level subsystems. Each subsystem is then refined in yet greater detail, sometimes in many additional subsystem levels, until the entire specification is reduced to base elements. However, black boxes may fail to elucidate elementary mechanisms or be detailed enough to realistically validate the model, giving no biological interpretability to the estimated parameters (e.g. sensitivity or behavior system).

However, we will classify the methods upon the type data needed for modeling:

- **Steady state methods:** Bottom-up methods that use steady state data to estimate the parameters, and can generally be defined as region restriction-based methods. These methods restrict the parameter space such that one does not have one particular parameter set but an ensemble of parameter sets. It quantifies the uncertainty of the “true” parameter set and gives a region of feasible parameter sets given the steady-state data.

- **Steady state and time course methods:** Methods that either use steady-state or time course data to estimate the parameter set. All of the methods optimize an objective function to find the most likely parameter set and can be classified as top-down approaches.

- **Time course methods:** These top-down methods need a limited number time-course data to estimate the parameters.
5.2 Steady state methods

5.2.1 Prediction of Enzyme Kinetic Parameters Based on Statistical Learning

This section describes a method constructed by Borger [32]. As already indicated the values of enzyme parameters are a key requisite for the kinetic modeling of biochemical systems. The method predicts parameters (e.g. Michaelis-Menten constant $K_m$) by a linear regression model in which the substrate, the combination enzyme-substrate and the combination organism-substrate have a linear effect on the logarithmic parameter value.

To do so they retrieve a set of measured Michaelis-Menten constants from databases. For each chosen metabolite, the data for all $K_M$ related to this metabolite is found and stored together with their associated enzymes (defined by its EC-number) and the organisms in which the values were measured. This results into a data triple ($K_M$ values, EC number, and organism). Next they construct a data matrix $X$, with rows corresponding to the EC-numbers and the columns correspond to the individual organisms, containing the logarithmic $K_M$ values. When enzymes participate in multiple reactions in the same organism, the average of the logarithmic parameters is taken. Some elements of the matrix shall stay empty as the $K_M$ value has not yet been measured or the enzyme has not (yet) been found in the genome sequence. The aim of the method is to predict these values based on the known elements of the matrix.

To this end, a linear regression model is defined:

$$(K_M)_{ij} = e^{\mu + \alpha_i + \beta_j + \epsilon_{ij}} \quad (5.1)$$

Taking the natural logarithm from (5.1):

$$x_{ij} = \mu + \alpha_i + \beta_j + \epsilon_{ij} \quad (5.2)$$

where $\mu$ is the general mean, $\alpha_i$ the effect of enzyme $i$, $\beta_j$ the effect of organism $j$ and $\epsilon_{ij}$ is identically independent distributed (i.i.d) Gaussian noise ($N(0, \sigma^2)$). Given the experimental data, we can compute the effects $\mu$, $\alpha_i$ and $\beta_j$ by a least squares fit minimizing the sum of quadratic residuals. The general estimate is:

$$x_{ij}^* = \mu + \alpha_i + \beta_j \quad (5.3)$$

but in cases in which data is missing, only a subset of $\alpha$, $\beta$ and $\mu$ can be used for estimation. Figure 5.1 illustrates this method of some logarithmic $K_M$ values of enzymes measured in 9 different organisms for one particular substrate. The empty spaces in the left matrix are values that yet have to be predicted and the right matrix shows the parameters after estimation. Note that if there are high values in a certain row or column it is likely that the predicted values in that row or column also have a high value.
Figure 5.1: Matrices before and after the application of the regression model. In the left matrix one can see the known logarithmic Michaelis-Menten constants before the regression. The grayscale bar denotes the magnitude of the values. In the right matrix one can see the Michaelis-Menten constants after regression.

To validate the method, the authors use leave-one-out cross-validation (LOOCV). The error between every known parameter and its prediction is computed and averaged, $\sigma_{\text{pred}} = \sqrt{\langle (x^* - \bar{x})^2 \rangle}$. Another measure the authors propose is the root mean square error of the original data $\sigma_x = \sqrt{\langle (x - \bar{x})^2 \rangle}$, addressing the uncertainty of the experimental values.

Figure 5.2 illustrates the results of the estimation procedures for four different metabolites. It shows that the linear model gives a better estimate of the parameter than the uncertainty of the data and that the true and predicted values have a moderate to high correlation.

Figure 5.2: Scatter plots of the predicted vs. the true Michaelis-Menten constants for four different metabolites (D-glucose ($\Delta$), Ethanol ($\times$), NADP+ ($+$) and Pyruvate ($^*$)). The error between the brackets is the root mean square error of the original data.
**Conclusion**

This method focuses on the estimation of these parameters from experimentally determined values stored in databases. The assumption is made that there is a direct correspondence for all parameter values that belong to the same enzyme and metabolite, but originate from different organism. There is also a biological interpretation in this estimation. The Michaelis-Menten constants in cells are an outcome of evolution: the $K_M$ values may be conserved between related species and most probably they have also been shaped by functional requirements. Both factors would lead to similarities between related Michaelis-Menten constants in different species. The use of this method is not limited to Michaelis-Menten constants, but can also be used for the maximal rate velocities, reaction rate coefficients and dissociation constants. These are all affected by the enzyme of the reaction, the organism in which the reaction takes place and the substrates of the reaction. Although the method is simplistic, one could use these predicted values as rough estimates for initial values in optimization procedures.

**Limitations**

**Multiple values**: If an enzyme participates in multiple reactions in an organism it may have multiple Michaelis-Menten constants which are averaged during constructing the data matrix. This could lead to inaccurate estimation of the parameters. An enzyme which participates in multiple reactions, would in this case have the same Michaelis-Menten constant for all reactions, which is not likely. A survey of the BRENDA database showed that this is indeed not the case.

**Validation**: The fact that the regression gives better estimates than a constant line does not imply that the regression is a guaranteed success. Statistics to see if the linear regression is the optimal type of regression (e.g. $R^2$ statistic) has not been shown and the results generally show that this regression does not give a large improvement. Furthermore, a correlation between the regression variables (organism, enzyme, etc.) does not imply causation and could limit the models capabilities by confounding variables.

**Evolution assumption**: The assumption that the kinetic parameters are preserved due to functional requirements, is likely untrue if the organisms in the regression have a large phylogenetic distance. Only simple organisms (prokaryotes) have an almost complete parameter set and should not be used for estimating parameters for complex organisms (eukaryotes).

**Noise handling**: The method does not address the integration of prior knowledge of measurement noise, but was one of the motivations addressed by the authors. All experimentally determined parameters can be assumed to be corrupted by noise, not modeling this fact could lead to severe limitations in the estimation.
5.2.2 Parameter space restriction

Introduction

Some methods currently used to estimate kinetic parameters do not give a specific parameter set, but restrict the parameter space to give an ensemble of parameter sets which can reproduce the experimental data. This restriction can be performed by hard constraints or can be given by parameter distributions. In this section we will discuss three methods that restrict the parameter space in such a way.

k-Cone Analysis

Enzyme assays inference kinetic parameters by isolating other enzymes and molecules that could interfere with the enzyme activity of particular reaction. These conditions are different from the in vivo conditions and are thus one of the mean reasons why in vivo derived kinetic models sometimes do not adequately describe in vivo phenotypes. To cope with this problem Famili [33] proposed a framework that integrates steady-state in vivo data with a constraint-based modeling approach, to construct a convex kinetic solution space (k-cone) which determines all candidate numerical values of reaction rate coefficients.

We begin by writing the system of ODEs in the mathematical convenient expression, seen in chapter 3 (3.1), and state that the system is in chemical equilibrium (3.2). This indicates that the vector of reaction rates must reside in the null space of the stoichiometric matrix. It is assumed that the enzymatic reactions can be modeled by the laws of mass action, giving the reaction rate vector:

\[ \mathbf{v} = \text{diag}(\mathbf{\bar{c}}) \mathbf{\bar{k}} \]  

(5.3)

where \( \mathbf{\bar{k}} \) is a vector with all reaction rate coefficients and \( \mathbf{\bar{c}} \) is a vector of numerical product of steady state substrate concentrations for each reaction, \( c_j = \prod_{i=1}^{p_j} x_i^{s_{ij}} \) (e.g. for \( 2x_1 + x_2 \rightarrow x_3 \), \( c_3 = x_1^2 x_2 \)). Note that if a reaction is reversible, it is decoupled into a forward and a backward reaction. The assumptions lead to:

\[ N\mathbf{v} = N\text{diag}(\mathbf{\bar{c}}) \mathbf{\bar{k}} = M\mathbf{\bar{k}} = 0 \]  

(5.4)

indicating that the vector \( \mathbf{\bar{k}} \) is in the null space of the matrix M. The basis vectors of this null space span all possible kinetic values that satisfy (5.4). The addition of stoichiometric constraints and the integration of equilibrium constants results in a convex cone in the parameter space. Figure 5.3 illustrates such a convex k-cone in the parameter space.
To calculate the relationship between kinetic parameters measured in vitro and the k-cone a combination of linear and nonlinear optimization methods are proposed:

Calculating the closest distance between a vector \( k' \) and the k-cone

Assuming we experimentally obtained kinetic parameters \( k' \), we would like to know the closest point in the k-cone feasible to reproduce the observed steady state \( \textit{in vivo} \) data. This results in a quadratic programming problem:

\[
\begin{align*}
\min_{k} & \| k - k' \|_2 \\
\text{Subject to:} & \ M^T k = 0 \quad (5.5) \\
& 0 \leq l_i \leq k_i \leq u_i, \quad i = 1, \ldots, n
\end{align*}
\]

where \( l \) and \( u \) are the lower and upper bounds of the individual parameters \( k_i \).

Calculating the closest distance from a vector \( k' \), given experimental error, to the k-cone

To model measurement error the following was proposed:

\[
N \text{diag}(C + \delta C) \tilde{k} = N \text{diag}(\tilde{C}) \tilde{k} + N \text{diag}(\delta \tilde{C}) \tilde{k} = M \tilde{k} + \delta \text{diag}(\delta \tilde{y}) \tilde{k}
\]

The minimization problem becomes:

\[
\min_{k, \delta} \| k - k' \|_2 \quad \text{Subject to:} \ M \tilde{k} + \delta \text{diag}(\delta \tilde{y}) \tilde{k} = 0 \quad (5.6)
\]

\[
\delta M = S \text{diag}(\delta \tilde{C})
\]

\[
-1 < \delta y_i < 1 \\
0 \leq l_i \leq k_i \leq u_i, \quad i = 1, \ldots, n
\]

The vector \( \delta \tilde{y} \) represents the extent of the error in each concentration and is used as an optimization variable to identify any point within the defined experimental error. It takes into account the entire space of concentrations within the experimental error.

Calculating the minimum required number of changes in \( k' \) to project it into the k-cone

This is a mixed-integer linear programming (MILP) problem:

\[
\begin{align*}
\min_{k, \theta} & \sum_{i=1}^{n} \theta_i \\
\text{Subject to:} & \ |k - k'| \leq N \tilde{\theta} \quad (5.7) \\
& M \tilde{k} = 0 \\
& 0 \leq l_i \leq k_i \leq u_i, \quad i = 1, \ldots, n
\end{align*}
\]

where \( \tilde{\theta} \) is the vector of Boolean variables, \( n \) is the dimension of the kinetic space and \( N \) is an arbitrary large positive number. If an in vitro kinetic constant is changed, than its corresponding Boolean variable \( \theta_i \) is set to 1. The solution of this optimization problem is a vector of kinetic constants in the k-cone one can obtain by the least number of changes in the in vitro measured kinetic parameters.
Human red blood cell

The kinetic model of the Human red blood cell is quite large constituting many reactions and parameters. The k-cone was constructed under four different psychological conditions all generating a different steady state vector of metabolite concentrations. For the human red blood cell data, it is very difficult to comprehend the restriction of the parameter space. The parameter space can be characterized by uniformly sampling the k-cone using a randomized Monte Carlo approach [35], by circumscribing it by a parallelepiped. By projecting the samples on the axes of the parameters we see how the k-cone is constructed. The histogram plot of the sample points illustrates the allowable range of kinetic variables and its height indicates the depth of the solution space at different values. Figure 5.4 shows this process.

![Figure 5.4: Uniform sampling of the parameter space.](image)

Furthermore, the experiment demonstrates how a network under different steady-state conditions may have concentration values that change significantly and result in different condition-dependent k-cones. Assuming the enzyme concentrations do not change, the actual parameter vector \( \mathbf{k} \) must satisfy all conditions simultaneously: we can use the intersection of the k-cones to restrict the parameter space even more as illustrated in Figure 5.5.

![Figure 5.5: Intersection of k-cones obtained from different experimental conditions](image)
Conclusion

This work has introduced a new approach for integrating steady state \textit{in vivo} data with constraint-based modeling to construct a kinetic solution space for large-scale metabolic networks. The range of feasible numerical values for the parameters can be reduced by computing the k-cone under different conditions. Using optimization algorithms we can recalculate \textit{in vitro} measured parameters such that they reproduce \textit{in vivo} phenotypes.

In addition to characterizing a feasible solution space for kinetic parameters in biological networks, k-cone analysis can be used to compute the kinetic solution space a priori to kinetic model development, thereby reducing time and effort involved in model building and parameter adjustments. Finally, the k-cone may give a (set of) initial vector(s) as starting points for top-down methods.

Limitations

\textbf{Simplification:} The enzymatic reactions in the metabolic network are modeled by laws of mass action, which implies that the phenomenon of enzyme saturation and allosteric regulation are not modeled. For some reactions this can be a severe simplification and could result in accurate predictions as the k-cone does not model reality.

\textbf{Enzyme concentration:} An assumption is that enzyme concentrations stay constant over time. In some experiments the enzyme concentration changes and thus is a simplification of the reality. To take into account the nonlinear nature of metabolic reactions, enzyme concentrations may have to be incorporated.

\textbf{Integration of prior knowledge on noise:} The authors propose to use a method that adds a concentration, $\delta C$, to the concentration vector which together with the noise vector $\delta y$ models the noise; but it is unclear how to measure or define $\delta C$.

Biochemical networks with uncertain parameters

Introduction

To estimate parameters \textit{Liebermeister} [36] quantified uncertain knowledge about parameters by probability distributions. As seen in Chapter 4 the dependent parameters can be expressed by thermodynamically independent variables which are characterized by these distributions. From these distributions we can calculate the distributions of model variables (e.g. flux). This probabilistic framework allows the study of metabolic correlations and it provides simple measures of variability and stochastic sensitivity.

Parameter distribution

The parameter distribution is characterized by an expected value, $\langle \ln(\tilde{p}) \rangle = \ln(\tilde{p}^0)$, and the covariance matrix $\text{cov}(\tilde{p})$ of the logarithmic values of the independent parameters. Furthermore it is assumed that the parameter values are log-normal distributed. To see if this assumption is valid, we conducted a survey. We extracted Michaelis-Menten constants and reaction rate coefficients from the BRENDA database and log-transformed them to see if the distributions followed an approximate log-normal distribution, shown in Figure 5.6. It is shown that the Michaelis-Menten constants can be fitted quite well with the log-normal distribution whereas the reaction rate coefficients have a less good fit. For now we assume that the parameters are distributed log-normal, but we shall discuss this assumption later in the limitations of this method.
As shown earlier the model is thermodynamically constrained, and the basic idea is to express the kinetic parameters (e.g. reaction rate constants) by other underlying parameters (e.g. equilibrium constants) until we reach a set of basic parameters that can be chosen independently. The dependencies among the parameters in this method are illustrated in Figure 5.7. An assumption made is that the parents constructing a child parameter are independent of each other and that the basic parameters (e.g. Gibbs free energy or enzyme activities) are either known or drawn from a fitted log-normal distribution. The dependence assumption can have a strong impact on the modeling results. Note that the independencies in this method are not meant to be biological facts, but modeling assumptions.

![Figure 5.7: Dependence graph for kinetic parameters](image)

The forward and reverse reaction rate coefficients for reaction $i$ can be expressed by their parents as:

$$k_i^+ = r_i u_i \sqrt{q_i} \quad (5.8)$$
$$k_i^- = \frac{r_i u_i}{\sqrt{q_i}} \quad (5.9)$$

where $q_i$ is the equilibrium constant, $u_i$ is the enzyme activity and $r_i$ is related to the free energy barrier of reaction $i$. The equations (5.8-5.9) are chosen such that $q_i = \frac{k_i^+}{k_i^-}$ and both reaction rate coefficients are proportional to the enzyme activity $u_i$. The kinetic parameters are constructed by the independent system parameters in a multiplicative fashion. This makes the logarithmic values additive.

$$\ln(k_i^+) = \ln(r_i) + \ln(u_i) + \frac{1}{2} \ln(q_i) \quad (5.10)$$
$$\ln(k_i^-) = \ln(r_i) + \ln(u_i) - \frac{1}{2} \ln(q_i) \quad (5.11)$$
In vector notation:

\[
\ln(k) = \ln(u) \begin{pmatrix} 1 \\ 1 \end{pmatrix} + \ln(r) \begin{pmatrix} 1 \\ 1 \end{pmatrix} + \frac{1}{2} \ln(q) \begin{pmatrix} 1 \\ -1 \end{pmatrix} \tag{5.12}
\]

Due to the independence assumption the expected value and the covariance matrix can easily be calculated from the independent parameters. If these logarithmic independent parameters are distributed normally than the dependent parameters are also distributed normally.

**Computing the parameter distribution**

Equation (5.12) shows the relationship between equilibrium constants, enzyme activities and kinetic parameters. For other rate laws the following form for this equation is proposed:

\[
\ln(k^{(i)}) = \tilde{a}^{(i)} \ln(u_i) + \tilde{b}^{(i)} \ln(q_i) + \tilde{c}^{(i)} + D^{(i)} \ln(\tilde{z}^{(i)}) \tag{5.13}
\]

where \(\ln(\tilde{z}^{(i)})\) is a vector with variables being i.i.d. standard normal (presumably noise) and \(\tilde{c}^{(i)}\) holding constants specific for the modeled rate law. By merging the vector and constructing block matrices we expand (5.13):

\[
\begin{pmatrix}
\ln(k^{(1)}) \\
\ln(k^{(2)})
\end{pmatrix} = \begin{pmatrix}
\tilde{a}^{(1)} & 0 \\
0 & \tilde{a}^{(2)}
\end{pmatrix} \begin{pmatrix}
 u_1 \\
 u_2
\end{pmatrix} + \begin{pmatrix}
\tilde{b}^{(1)} & 0 \\
0 & \tilde{b}^{(2)}
\end{pmatrix} \begin{pmatrix}
 q_1 \\
 q_2
\end{pmatrix} + \begin{pmatrix}
\tilde{c}^{(1)} & 0 \\
0 & \tilde{c}^{(2)}
\end{pmatrix} + \begin{pmatrix}
 D^{(1)} & 0 \\
0 & D^{(2)}
\end{pmatrix} \begin{pmatrix}
\ln(\tilde{z}^{(1)}) \\
\ln(\tilde{z}^{(2)})
\end{pmatrix} \tag{5.14}
\]

Calculating the expected value and covariance matrix for the model parameter vector:

\[
< \ln(k) > = A < \ln(u) > + B < \ln(q) > + \tilde{c} \tag{5.15}
\]

\[
\text{cov}(\ln(k)) = A \text{cov}(\ln(u)) A^T + B \text{cov}(\ln(q)) + D D^T \tag{5.16}
\]

It shows that it easily propagates uncertainty from independent basic parameters to all model parameters.

**Distribution of model variables**

To calculate the posterior distribution of the dependent parameters using steady-state data we need to know the distribution of the model variables. If the parameter distribution is narrow enough we can approximate the variable distribution by a first-order Taylor expansion around \(\bar{p}^0\). Figure 5.8 illustrates the linear expansion for one parameter \(p\) and one variable \(y\). If \(\ln(y)\) is linearized around \(\bar{p}^0\), then \(\ln(y)\) is also normal with a standard deviation of \(R\), where \(R\) is the slope of the linearized line.
The slope of this line is defined by the normalized response coefficients from MCA (supplement, section S3.3) and is defined as:

$$R_{ik} = \frac{\partial \ln y_i}{\partial \ln p_k} = \frac{p_k \frac{\partial y_i(\vec{p})}{\partial p_k}}{y_i} \tag{5.17}$$

$$R_{ikl}^{(2)} = \frac{\partial^2 \ln y_i}{\partial \ln p_k \partial \ln p_l} = \frac{p_k p_l \frac{\partial^2 y_i(\vec{p})}{\partial p_k \partial p_l}}{y_i} + R_{ik} \delta_{ik} - R_{ik} R_{il} \tag{5.18}$$

where Kronecker’s delta $\delta_{ik}$ denotes the elements of the identity matrix when $i = k$. Using a second order Taylor expansion we get:

$$\ln y_i(\ln \vec{p}^0 + \Delta \ln \vec{p}) \approx \ln y_i(\ln \vec{p}^0) + \sum_m R_{im} \Delta \ln p_m + \frac{1}{2} \sum_{m,n} R_{imn}^{(2)} \Delta \ln p_m \Delta \ln p_n + O(\Delta \ln \vec{p}^2) \tag{5.19}$$

Under the assumption that the functions $\vec{y}(\vec{p})$ are continuous, positive and differentiable we calculate the normal distribution of the variables. Using (5.19) and the assumption that $\Delta \ln p_m = 0$ and $\Delta \ln y_m = R \Delta \ln p_m$:

$$< \ln y_i > = \ln y(\ln \vec{p}^0) \tag{5.20}$$

$$\text{cov}(\ln \vec{y}) = R \text{cov}(\ln \vec{p}) R^T \tag{5.21}$$

**Qualitative variables**

We can also use this method to make statements about the variables. Let us assume that we have a forward and backward flux velocity $y_1$ and $y_2$. If the vector $\ln \vec{y} = (\ln y_1, \ln y_2)^T$ follows a bivariate normal distribution with mean $< \ln \vec{y} >$ and covariance matrix $\text{cov}(\ln \vec{y})$, the probability for a positive total flux is given by:

$$P(y_1 > y_2) = \phi \left( \frac{< \ln y_1 > - < \ln y_2 >}{\sqrt{\text{var}(\ln y_1) + \text{var}(\ln y_2) - 2 \text{cov}(\ln y_1, \ln y_2)}} \right) \tag{5.22}$$
Posterior distribution

The parameter distributions can be used as prior distributions for Bayesian parameter estimation. If we would collect experimental data for some of the model variables we can estimate the posterior distribution of the model parameters. Let $\hat{y}^{\text{exp}}$ denote the measurement vector. We assume that the experimental data has some multiplicative noise. This is modeled by:

$$\ln \hat{y}^{\text{exp}} = \ln \bar{y} + \sigma \hat{n} \quad (5.23)$$

where $\hat{n}$ is a vector with independent standard Gaussian random variables. If the variables are linearized around the reference parameters, we obtain an approximation for the posterior of $\ln \hat{p}$ by:

$$< \ln \hat{p} >_{\text{post}} = (\sigma^{-2} R^T R + C^{-1})^{-1}(\sigma^{-2} R^T \bar{w} + C^{-1} \ln \hat{p}^0) \quad (5.24)$$

$$\text{cov}_{\text{post}}(\ln \hat{p}) = (\sigma^{-2} R^T R + C^{-1})^{-1} \quad (5.25)$$

where $\bar{w} = \ln \hat{y}^{\text{exp}} - \ln \bar{y}(\ln \hat{p}^0) + R \ln \hat{p}^0$.

Conclusion

Liebermeister constructed a method where the model parameters are composed of independent system parameters which satisfy the thermodynamic constraints. With the assumption of log-normal distribution of these system parameters we can easily construct the distribution of the model parameters. From these distributions we can make statements about the dynamical properties of the model. It is possible to integrate information from experimental data into the distribution of the model parameters, as it is hard to estimate the true parameters, but by using distributions we can assert the uncertainty of the estimation.

One of the downsides of this paper is that the authors do not show if the integration of experimental data leads to a major improvement of the estimation, such as narrower distributions.

Limitations

**Relationship:** The model parameters are constructed from independent system parameters. With mass-action kinetics the reaction rate coefficients can be constructed from the enzyme activity, equilibrium constant and a proportionality constant. Although it seems logical that these parameters are independent it remains unclear how a product of this constants expresses the reaction rate coefficients.

**Rate laws:** It is unclear how we should formulate (5.14) for other rate laws. The authors give additional formulas for common rate laws (e.g. Michaelis-Menten kinetics), which are based on assumptions contradicting the rate laws or physics (e.g. the Michaelis-Menten constant is independent of the maximal velocity). Such assumptions could lead to severe problems.

**Log-normal distribution:** It is assumed that the model and system parameters are log-normal distributed. Figure 5.9 illustrates that not all parameter distributions can be approximated by a log-normal distribution as already shown in this section.

![Figure 5.9: Distribution of natural logarithmic transformed parameters](image)
**Linearization:** The question is how well the method performs if the function $\ln \bar{y}$ is highly nonlinear, as it can have significant effects on the distribution estimation. In rare cases the function of the model variables can be non-linear which cannot be linearized well.

**Distribution:** The linearization of the model variable function can only be performed in the parameter distribution is narrow enough. In the estimation of the prior parameter distributions one would normally estimate the mean value and covariance for each type of parameter from literature. This implies that parameters of the same type have the same mean value and variance, resulting in a spherical distribution. We would then obtain a parameter distribution which is not narrow and thus could have a significant negative impact on the linearization.

*Bringing metabolic networks to life: integration of kinetic, metabolic and proteomic data*

**Introduction**

This method tries to calculate a posterior distribution of the parameters values within a Bayesian framework based on the integration of kinetic, thermodynamic, metabolic and proteomic data. It is assumes that the structure of the metabolic network is known and that the reactions are modeled by convenience kinetics. As data is sparse or unreliable, we do not describe the model parameters by sharp values, but by a joint posterior distribution from which we can also see the correlation between parameters. The posterior distribution summarizes all information that has been put into the model and can be used to provide parameter ranges for further modeling. First we extract the independent system parameters, from databases and calculate their mean value and uncertainty, quantifying this in the prior distribution $p(\theta)$. The next step is the incorporation of the information about the enzyme kinetics. The last step is the integration of metabolic information into the posterior distribution. To improve the fit for the joint posterior distribution we can integrate metabolic data from several experiments. Figure 5.10 illustrates the integration pipeline and different types of data.

---

**Figure 5.10:** Data integration pipeline. (A) Metabolic network is translated into a kinetic model. (B) The model parameters for the enzyme mechanism are used to obtain a first, kinetics-based distribution for the enzyme parameters (D). A fit to the metabolic steady state data, gives a second metabolic-based distribution of the system parameters (E).
Methods

The vector $\hat{\theta}^{kin}$ contains the logarithmic independent parameters from convenience kinetics and can be extracted from literature. The vector $\hat{x}^*$ contains the dependent parameters, which should be experimentally inferred or extracted from databases. To model cells in specific experimental situations, we specify additional steady state parameters: a specific steady state $m$ is characterized by enzyme concentrations $h_i^{(m)}$, where $i$ is the index of the reactions, and fixed concentrations $s_i^{(m)}$ for external metabolites. We collect their logarithmic values in the vector $\hat{\theta}^{met}$. For convenience we define the parameter vector $\hat{\theta} = (\hat{\theta}^{kin}, \hat{\theta}^{met})$. First the system parameters are fitted to thermodynamic and kinetic data, such. Using convenience kinetics, the vector $\hat{x}$ can be calculated from the vector $\hat{\theta}^{kin}$, through:

$$\hat{x}(\hat{\theta}^{kin}) = R_x^* \hat{\theta}^{kin} \quad (5.26)$$

By augmenting the transformation matrix $R_x^*$ with zero vectors we can use the complete parameter vector in (5.26) instead of $\hat{\theta}^{kin}$. In the second step the parameter estimates are further improved by a fit to metabolic data from one or more steady states, collected in the vector $\hat{y}^*$. What we want to know is the distribution of the system parameters based on the given experimental data $\hat{x}^*$ and $\hat{y}^*$, expressed by:

$$p(\hat{\theta}|\hat{x}^*, \hat{y}^*) \sim p(\hat{x}^*, \hat{y}^*|\hat{\theta})p(\hat{\theta}) \quad (5.27)$$

Under the assumption that the experimental data of $\hat{x}^*$ is independent of the experimental data, $\hat{y}^*$, we can state (5.27) as:

$$p(\hat{\theta}|\hat{x}^*, \hat{y}^*) \sim p(\hat{x}^*|\hat{\theta})p(\hat{\theta}) = p(\hat{y}^*|\hat{\theta})p(\hat{x}^*|\hat{\theta})p(\hat{\theta}) \quad (5.28)$$

It is assumed that the prior distribution of $\hat{\theta}$ is a multivariate Gaussian distribution:

$$p(\hat{\theta}) \sim N(\bar{\theta}(0), C(0)) \quad (5.29)$$

Mean and variance of this distribution are chosen for each system parameter depending on its type, such as equilibrium constant. These statistics can be derived from empirical distributions of parameter values extracted from databases. It is assumed that the system parameters are uncorrelated resulting in a diagonal covariance matrix $C(0)$. The likelihood functions $p(\hat{x}^*|\hat{\theta})$ and $p(\hat{y}^*|\hat{\theta})$ represents a simple model of the measurement process. We assume that the predicted values $\hat{x}^*$ and $\hat{y}^*$ equal the values predicted by a model plus some uncorrelated additive Gaussian noise:

$$p(\hat{x}^*|\hat{\theta}) \sim N(\bar{x}(\hat{\theta}), C_x) \quad (5.30)$$
$$p(\hat{y}^*|\hat{\theta}) \sim N(\bar{y}(\hat{\theta}), C_y) \quad (5.31)$$

This noise assumption implies that the covariance matrices can be expressed as $C_x = diag(\bar{\sigma}_x)$ and $C_y = diag(\bar{\sigma}_y)$, where the vectors $\bar{\sigma}_x$ and $\bar{\sigma}_y$ contain the noise level for each measurement. The kinetic parameter vector $\bar{x}(\hat{\theta})$ can be calculated with convenience kinetics, but the metabolite concentrations and fluxes contained in $\bar{y}(\hat{\theta})$ are computed numerically by solving steady state equations.

Theoretically we can obtain the posterior distribution by simply inserting (5.29-5.31) into (5.28), but this does not show how we can actually compute it. Sampling methods are unfeasible with such a large
number of model parameters. The only solution is to approximate the posterior by a Gaussian distribution around a local maximum, the so-called posterior mode. Instead of using \( p(\theta | \mathbf{x}^*, \mathbf{y}^*) \) self, let us use the following function.

\[
F(\theta) = -2 \ln p(\theta | \mathbf{x}^*, \mathbf{y}^*) = (\hat{\theta} - \hat{\theta}_{(0)})^T C_{(0)}^{-1} (\hat{\theta} - \hat{\theta}_{(0)}) + (\hat{x}^* - \hat{x}(\theta))^T C_x^{-1} (\hat{x}^* - \hat{x}(\theta)) + (\hat{y}^* - \hat{y}(\theta))^T C_y^{-1} (\hat{y}^* - \hat{y}(\theta)) \tag{5.32}
\]

As long as \( F(\theta) \) is quadratic in \( \theta \) the posterior is a Gaussian distribution. This is indeed the case for (5.32) as long as we do not consider the metabolic data \( \mathbf{y}^* \). The first two terms are quadratic in \( \theta \), so we can use them to approximate the first posterior, called the kinetics-based posterior.

**Kinetics-based posterior**

The posterior probability density in this case reads \( p(\theta | \mathbf{x}^*) \sim P(\mathbf{x}^* | \theta) p(\theta) \) and has a multivariate normal distribution \( N(\tilde{\theta}_{(1)}, C_{(1)}) \). The derivation of the mean and covariance matrix is extremely long and can be found in the supplement section S5.1, resulting in:

\[
\tilde{\theta}_{(1)} = (C_0^{-1} + (R_0^x)^T C_x^{-1} R_0^x)^{-1} (C_0^{-1} \tilde{\theta}_0 + (R_0^x)^T C_x^{-1} \mathbf{x}^*) \tag{5.33}
\]

\[
C_{(1)} = (C_0^{-1} + (R_0^y)^T C_y^{-1} R_0^y)^{-1} \tag{5.34}
\]

**Metabolic based posterior**

The likelihood \( p(\mathbf{y}^* | \theta) \) is hard to compute as \( \mathbf{y}(\theta) \) depends non-linearly on the system parameters. To cope with this problem we take a first-order Taylor expansion of this function around the expansion vector \( \theta \).

\[
\hat{y}(\theta) \approx \mathbf{y}(\theta) + R_0^y (\theta - \hat{\theta}) \tag{5.35}
\]

where the matrix \( R_0^y \) contains the sensitivities \( R_0^{yi} = \frac{\partial y_i}{\partial \theta_m} \). This posterior of the linearized model is a multivariate Gaussian distribution \( N(\tilde{\theta}_{(2)}, C_{(2)}) \). The derivation of the mean values and the covariance matrix is too long to show here and can be found in the supplement section S5.1, resulting in:

\[
\tilde{\theta}_{(2)} = \theta + (C_1^{-1} + (R_0^y)^T C_y^{-1} R_0^y)^{-1} (C_1^{-1} (\tilde{\theta}_1 - \theta) + (R_0^y)^T C_y^{-1} (\mathbf{y}^* - \mathbf{y}(\theta))) \tag{5.36}
\]

\[
C_{(2)} = (C_1^{-1} + (R_0^y)^T C_y^{-1} R_0^y)^{-1} \tag{5.37}
\]

Although we have the formulas, we still do not have the expansion vector \( \theta \). For this expansion vector we would like to take the maximum of the posterior. This algorithm starts with parameters and initial conditions that correspond to a chemical equilibrium. We linearise the model, compute the posterior mode and move the current parameter vector towards this mode using a relaxation constant \( \lambda \). We then compute the posterior (5.44-5.45) resulting from the linearized model and move our expansion point towards the parameter set \( \tilde{\theta}_{(2)} \). This procedure is iterated until convergence. In the end we have the expansion point and the matrix of sensitivities and we can calculate the final posterior distribution.
Results

Chassagnole [38] presented a detailed kinetic model for the threonine biosynthesis pathway and experimentally determined a large number of system parameters. To test the framework other independent system parameters were calculated from the molecular structures of the metabolites and the energy constants \( k^p \) were determined by the known reaction rate coefficients. To create artificial experimental data for the kinetic parameters Gaussian noise was added to the model parameters after transforming the original system parameters. To create artificial metabolic data, the authors randomly chose concentrations for the fixed metabolites and the enzymes by multiplying the experimentally determined values with independent log-normal random numbers. From this point on the steady states for five such random assignments were computed and was repeated ten times. The resulting steady state quantities where used as artificial metabolic data. Figure 45 illustrates the estimated distributions for the threonine model.

![Figure 5.11: Distributions in the threonine model. Left: The prior and kinetics-based posterior in the threonine model. On the horizontal axis one can see the system parameters of this model while on the vertical axis we can see the range of the logarithmic values. Black (■): True parameter values. Bars indicate the marginal distribution of a system parameter with one standard deviation from the mean. Light blue (○): prior distribution of the logarithmic parameters. Red (+): Likelihood function of the experimental kinetic values. Blue (*): Kinetics-based posterior. Right: Black (■): True parameter values. Blue (*): Kinetics-based posterior. Purple (◊): Metabolic-based posterior.](image)

The left image of Figure 5.11 shows that the integration of kinetic narrows the posterior distribution of the system parameters with respect to the prior distribution. The integration of metabolic data further narrows the marginal posterior distribution (metabolic-based posterior), as shown in the right image of Figure 5.11, but to a small extent. It does not show that the correlation between the parameters has become stronger and this can only be shown by an eigenanalysis of the covariance matrices, as shown in Figure 5.12. The left image shows the ordered eigenvalues of the covariance matrices and clearly illustrates that the magnitude of the eigenvalues for the kinetics-based posterior decreases significantly with respect to the prior distribution, resulting in a decrease of uncertainty. It also shows that some eigenvalues decrease for the metabolic-based posterior with respect to the kinetics-based posterior, but only slightly. For the kinetics-based posterior, the middle image clearly shows that most parameters are not correlated. Whereas for the metabolics-based posterior the right image clearly shows more and stronger parameter correlations, implying the integration of different kind of data give more informative distributions. If we for example know a subset of parameters this can generally give much information about the unknown but correlated parameters.
Figure 5.12: Eigenanalysis of the covariance matrices. Left: The eigenvalues, ordered in magnitude, of the three distributions. Middle: Eigenanalysis of the kinetics-based posterior. A row in this matrix corresponds to an eigenvalue with its associated eigenvector. The colored elements of the matrix correspond to the degree of participation for a parameter (column) to a eigenvector (row). Right: Eigenanalysis of the metabolic-based posterior.

Simulation

To test the parameter distributions in terms of predictive behavioral power, a simulation of the threonine model was performed by sampling parameter sets from the prior, first posterior and second posterior and look at their temporal behavior with respect to artificial time-course data. To assess how the time courses are distributed the simulation was performed 100 times for each distribution. After 5 minutes the concentration of the external metabolite aspartate was increased by a factor of 50. Figure 5.13 shows the results of the simulations for two different metabolites. The areas in these plots represent the concentrations obtained during the simulation. It clearly shows that the refined parameter distributions lead to better predictions of the dynamic behavior.

Conclusion

The method described in this section can be classified as a bottom-up procedure as we are piecing together different kinds of information to estimate the parameters. The posterior distribution represents the typical ranges of model parameters gained from integrating experimental data. This method does not assign one specific parameter set but quantifies the uncertainty about them through distributions. Data sources with a small uncertainty will have the greatest impact on the estimation and new data can recursively be integrated into the posterior. The accuracy of the predictions has been shown to increase considerably between the kinetics-based and metabolic-based posterior. Hence, the fit to metabolic data
adds important information to the parameter distribution and is contained in the parameter correlations rather than decreasing the uncertainty of the marginal distributions. As will be discussed later some top-down modeling methods will only try to optimize parameters such as reaction rate coefficients without quantifying uncertainty about them. These methods need an initial vector to start with and we can make use of the distributions estimated with this method.

**Limitations**

**Independence:** In the Bayesian framework the experimental data of the enzyme parameters is assumed to be independent of the metabolic data, which might hold untrue. If one were to change a sensitive reaction rate coefficient it could happen that the system goes to another steady state, yielding different steady state metabolite concentrations, indicating dependence.

**Transformation assumption:** It is assumed that the transformation of the system parameters to model parameters is close to the experimentally determined model parameters. First of all convenience kinetics was constructed to estimate the model parameters with independent parameters, such that we can tune them without violating any constraints (4.31). However, the paper of neither convenience kinetic nor this paper show that if one has the independent system parameters from literature its transformation is close to the true dependent parameter values with some additive noise (5.30). Such an assumption should be validated, but there seems to be no proof for it to be true.

**Prior knowledge:** In the results the authors have information normally not known during experimentation. They construct a set independent of system parameters and transform them to the enzyme kinetic parameters, \( \hat{x}(\hat{\theta}) \). To this vector they add some Gaussian noise to construct the vector \( \hat{x}^* \) of artificial experimental data. In this case we know that the calculated kinetic parameters and the artificial noisy experimental data are close to each other as they are constructed from the same set of system parameters (see previous point). Furthermore we exactly know what type and magnitude of noise was added to the vector such that we can express the uncertainty exactly. The same procedure of data generation has been performed for artificial metabolic data.

**Noise assumption:** It is assumed that the experimental and measurement noise are log-normal distributed. There is no proof or guarantee that is actually the case. If the noise is characterized by a different distribution this could lead to a sub-optimal posterior distribution, not modeling the correct natural phenomenon.
5.3 Optimization

5.3.1 Introduction

This section discusses the estimation of the model parameters by means of optimizing an objective function based upon steady state and time course data. Although this is common to all optimization methods for parameter estimation, it is possible to choose between different algorithms for optimization and different representations of the biochemical system. Also, additional statistics can be derived which make statements about the parameter space to be optimized and whether the experimental data has the capability of identifying each parameter uniquely.

5.3.2 Optimization of S-system models

Introduction

This method is developed by Gonzalez [39] and it changes the biochemical representation of the system equations into S-systems developed by Savageau [40-41] and further explained by Torres, Voit [42]. This system is a representation constructed by a non-linear approximation of the system equations and is further explained in the supplement, section S5.2. With this representation, the change of one metabolite over time is:

\[
\frac{dX_i}{dt} = V^+(X_1, \ldots, X_n, X_{n+1}, \ldots, X_{n+m}) - V^-(X_1, \ldots, X_n, X_{n+1}, \ldots, X_{n+m})
\]

where \(V^+\) represents the forward reaction rate (the construction of the metabolite) and \(V^-\) represents the reverse reaction rate (degradation metabolite). \(X_i\) from 1 to \(n\) represent the internal metabolite concentrations and from \(n + 1\) to \(n + m\) represents the external metabolite concentrations. We express the biochemical system with S-system models, reducing the model representation, as:

\[
\frac{dX_i}{dt} = \alpha_i \prod_{j=1}^{n+m} X_j^{g_{ij}} - \beta_i \prod_{j=1}^{n+m} X_j^{h_{ij}} \quad \text{for} \quad i = 1, 2, \ldots, n
\]

where \(g_{ij}\) and \(h_{ij}\) are called the kinetic order of metabolite \(j\) in reaction \(i\). \(\alpha_i\) and \(\beta_i\) are the rate constants.

Methods

From the stoichiometric matrix we already have a hunch which metabolites have an influence on the concentration of a specific metabolite. This can be reflected by the kinetic orders in the formulation. If a certain metabolite does not influence the dynamics of another metabolite than we can set its kinetic order equal to zero effectively dropping from (5.39). With the reduced formulation it is now the goal to determine the parameters that are left free with optimization. For this procedure the global optimization technique Simulated Annealing was chosen and has two major advantages:

1. Lower quality solutions can be accepted with a certain probability depending on the pseudo-temperature variable \(T\).
2. It functions much like a global optimization technique at high temperatures and implicitly switches to a local optimization behavior at lower temperatures.

We optimize the parameters through an objective function (5.40) such that the model outputs resemble the experimental time-course data as good as possible. If we have a large-scale model, the parameter space is
most likely multimodal, thus reducing the possibility to find the global optimum. This is the reason for choosing a global optimization algorithm as it in theory can find the global optimum. The residual sum of squares objective function is given by:

\[ E = \frac{\sum_{t=1}^{n} \sum_{i=1}^{m} (\hat{X}_{it} - X_{it})^2}{n \cdot m} \] (5.40)

where \( \hat{X}_{it} \) stands for the predicted value for constituent \( X_i \) at time point \( t \) and \( X_{it} \) is the experimentally observed value for the constituent \( X_i \) at time point \( t \). Figure 5.14 illustrates the pseudo-code for this Simulated Annealing procedure. Vector \( \vec{h} \) holds all the parameter values and is changed by the perturbation function \( p + k \cdot \ln \sqrt{E + 1} \cdot N(\bar{x}, \sigma^2) \).

```
1  initialize \( h \)
2  ctr ← 0
3  \( T ← T_0 \)
4  while \( T > T_{\text{min}} \)
5    for \( ctr ← 1 \) to \( ctr_{\text{max}} \)
6      \( h' ← h \)
7      for each kinetic parameter \( p \) of \( h' \)
8        \( p ← p + k \cdot \ln(\sqrt{E + 1}) \cdot N(\bar{x}, \sigma) \)
9        \( \Delta E ← \text{Error}(h') - \text{Error}(h) \)
10       if \( \Delta E ≤ 0 \)
11          \( h ← h' \)
12        else
13          \( h ← h' \) with probability \( \frac{\Delta E}{E} \)
14      end for
15  lower \( T \)
16  end while
```

Figure 5.14: Pseudo code for Simulated Annealing

The pseudo-code shows that some solutions are accepted based on the parameter \( T \) (pseudo-temperature) which decreases in time. The derivation and some statements about the perturbation function can be found in the supplement section S5.3. The reader is not obliged to read it but should make the section about the limitations more understandable. With respect to previous methods optimization can process time series data.

**Top-down modeling and pruning**

The method as just discussed is classified as a top-down approach. Closely related to this inverse problem is the task of identifying network structure. S-systems are well suited for structure identification problems because of their regular structure and the fact that each parameter has a well defined biological interpretation. Kinetic order defines the influence of a metabolite on the dynamics of another metabolite. One of the pruning steps used is removing constituents from the formula with significant small kinetic order and performing another round of optimization with this constraint to get a new solution. Figure 5.15 illustrates the iterative method together with the pruning steps to find the solutions.
One has to be very careful with the pruning steps as they may violate known properties of the model. Some guidelines to prevent these issues are:

- **Perform targeted experiments**: Lab experiments especially targeted at testing a particular pruning step can give more information.
- **Consistency with known biochemical facts**: Pruning steps should not violate any known property of the system being modeled.

### Results

The framework was tested on the cadBA system in *Escherichia coli*; system parameters were extracted from the literature. Figure 5.16 illustrates the reaction scheme of this model together with the system of ODEs defined by S-systems. From this system of ODEs artificial (noise-free) data was generated with the initial conditions show in Figure 5.16 and a measurement interval of 0.1 seconds.

A number of 18 free parameters need to be optimized. The rate constants were randomly assigned a value in the range [0, 90] and the kinetic orders in the range [-2, 2]. As was expected the initial guesses were far from the true solution with respect to the residual error and parameter values. Figure 5.17 illustrates that the time series of the estimated parameters follows closely those created from the true parameters.
Conclusion

We have seen that with the S-system representation and Simulated Annealing we can find the parameter values in a top-down approach. The results section shows that the parameter values of the optimization were very close to the true values leading to a small residual error. Using this method one can use steady state data as well as time series to find the parameters. One additional advantage is that the use of S-systems and pruning leads to a reduction in parameters.

Limitations

**Biological interpretability:** Optimization procedures, which can be classified as top-down approaches, are limited by the fact the estimated parameters can generally not be interpreted. It can, for example, not be stated if some parameters are sensitive or that a parameter is estimated with great uncertainty. In this case we want to control the model and not gain much knowledge about it.

**Parameter space:** The parameter space is very multimodal which could result in the optimization algorithm getting stuck in a local optimum. Using the global optimization method Simulated Annealing decreases this probability, but it does not guarantee that it will find the global optimum in finite time. In such a high-dimensional and complex search space, the algorithm will take a long time to complete and find a good solution; this is a limitation if time-constraints play a role.

**Measurement interval:** Optimization approaches can model time series data, but this method is applied on artificially generated time series with a time interval of 0.1 seconds. This is not (yet) possible as the fastest measurement intervals are currently in the order of a few seconds.

**Prior knowledge of measurement noise:** Prior knowledge of the measurement noise was not modeled in this method. The optimization procedure will try to fit the noisy measurements, resulting in inaccurate model parameters. In the results section the authors generated noise-free artificial data, which is not realistic, as most measured variables in the laboratory have somewhat around 30% measurement noise.

**Perturbation function:** The perturbation function of the Simulated Annealing procedure is based upon small measurement intervals, as already stated this is not the reality; this might lead to problems if one wants to model real data.
5.3.3 System identification through optimization

Introduction

This section shall be a summarization of three papers concerned with system identification through optimization. As these papers follow a chronological order in the development of the optimization techniques we shall not describe each paper separately, but combine them in one section.

Common methodology

The papers have in common that they are trying to optimize an objective function by tweaking the parameter values. Instead of the approximation by S-systems these methods make use of the more standard network equations as described in previous chapters.

First we state the objective function:

\[
J(p) = \int_0^t (\hat{y} - \hat{y}(x, p, t))^T W(t) (\hat{y} - \hat{y}(x, p, t)) \, dt \quad (5.41)
\]

Subject to: \(x(t_0) = x_0\) and \(p^l \leq p \leq p^u\)

As we do not have continuous measurements we have to discretize (5.41):

\[
J(p) = \sum_{i=1}^m \sum_{j=1}^n w_{ij} (y_{ij} - \hat{y}_{ij}(p, t))^2 \quad (5.42)
\]

where \(w_{ij} = 1/(\max_j \hat{y}_{ij}(p, t))^2\) is used to normalize the contributions to the cost function. These weights are chosen such that under the assumption that the measurement errors are i.i.d. normally distributed with zero mean and constant variance this function is a maximum likelihood estimator, and also called the weighted least squares estimator.

Non-linear optimization

To optimize the objective function we need to find out which optimization algorithm can be used best. Because of the nonlinear and constrained nature of biochemical systems we need to make use of nonlinear optimization. To know which types of non-linear optimization methods work best Mendes and Kell [43] tested different kinds to see which suited this problem best. The methods where classified into two categories of functions; local optimizers (e.g. Steepest descent) and global optimizers (e.g. Evolutionary programming). Some methods have additional advantages as they can handle constraints and boundaries. As expected the global optimization methods performed much better than the local optimization methods (data not show). This follows logically from the fact that the problem is very multimodal, so local optimizers likely find only local optima. Local optimizers will only finds the global optimum if it is in its basin of attraction.

To further test the methods a number of rate constants of the mechanism of irreversible inhibition of HIV proteinase are estimated. Data from five time courses at four different inhibitor concentrations measured fluorimetrically were used. The left image of Figure 5.18 illustrates the reaction mechanism of this process. We are trying to estimate the rate constants \(k_{22}, k_3, k_{42}, k_{52}\) and \(k_6\). The remaining rate constants are identified as known and shall not be optimized. The initial values for the optimized rate constants were extracted from the literature and it was assumed that there is a certain degree of uncertainty in the
values of the initial concentrations of substrate and enzyme and that the baseline of the fluorimeter was not exactly zero. Therefore we have 20 adjustable parameters: 5 rate constants, 5 initial concentrations of enzyme, 5 initial concentrations of the substrate and 5 offset values. Simulated Annealing yielded the best results with respect to equation (5.42) and it is not surprising that it is a global stochastic optimizer. Figure 5.18 illustrates the temporal concentration curves fitted on the artificial (noisy) experimental data generated with parameters found with Simulated Annealing (SA).

\[
\begin{align*}
M + M & \rightleftharpoons E & \quad k_{11} & \quad k_{12} \\
S + E & \rightleftharpoons ES & \quad k_{21} & \quad k_{22} \\
ES & \rightarrow E + P & \quad k_3 \\
E + P & \rightleftharpoons EP & \quad k_{41} & \quad k_{42} \\
E + I & \rightleftharpoons EI & \quad k_{51} & \quad k_{52} \\
EI & \rightarrow EJ & \quad k_6
\end{align*}
\]

Figure 5.18: Left: Reaction mechanism of irreversible inhibition of HIV proteinase. Right: Simulated smooth curves from the parameters obtained with SA together with the experimental data (noisy curves).

Global optimization methods:

With the evidence that global optimization methods perform better than local optimization methods Moles [44] tried to find out which type of global optimization method would generally give the best results. The use of several state-of-the-art deterministic and stochastic global optimization methods was explored. Stochastic methods ultimately rely on probabilistic approaches, given that random elements are involved these methods only have weak theoretical guarantees of convergence to the global optimum. Deterministic methods are those that provide a level of assurance that the global optimum will be found, however note that such algorithms cannot solve global optimum problems within finite time. Some deterministic methods have sound theoretical properties, but associated computational effort increases very rapidly, sometimes exponentially, with the problem size. Stochastic methods can locate the vicinity of the global solution with relative efficiency in modest computation time. If one would be satisfied by a solution that is less accurate then these methods are optimal as they are also easy to implement.

Figure 5.19: The metabolic pathway for this study. Solid arrows represent mass flow, dashed arrows kinetic regulation; arrows end represent activation, blunt ends inhibition. S and P are the pathway substrate and product and are held at constant concentration; M1 and M2 are intermediate metabolites of the pathway; E1, E2 and E3 are the enzymes; G1, G2 and G3 are the mRNA species of the enzymes.

Figure 5.19 illustrates the metabolic pathway used in this study on which to perform the optimization procedure. From the 8 ordinary differential equations we need to estimate 36 parameters. For the substrate
and the product there are four different concentration levels, as shown in Figure 5.20, from which 16 time series were generated (there are 16 combinations of substrate and product concentration levels.

![Figure 5.20: Substrate and product concentrations used to generate data. Bottom: Parameters used as true parameters.](image)

Table 5.1 shows all the different stochastic and deterministic optimizers used for estimation the parameters. It becomes clear that for this problem the Evolutionary Strategies are the best optimizers. Although they need a lot of simulations and take a lot of computation time the cost function minimization is almost optimal.

<table>
<thead>
<tr>
<th>Name</th>
<th>Description</th>
<th>Cost function</th>
<th>Simulations</th>
<th>CPU time (h)</th>
</tr>
</thead>
<tbody>
<tr>
<td>SRES</td>
<td>Evolution strategy with stochastic ranking</td>
<td>0.0013</td>
<td>28e5</td>
<td>39.42</td>
</tr>
<tr>
<td>uES</td>
<td>Unconstrained evolution strategy</td>
<td>0.0109</td>
<td>28e5</td>
<td>41.27</td>
</tr>
<tr>
<td>CMA-ES</td>
<td>Evolution strategy with covariance matrix adaption</td>
<td>37.881</td>
<td>756135</td>
<td>17.31</td>
</tr>
<tr>
<td>DE</td>
<td>Differential Evolution</td>
<td>151.779</td>
<td>22.5e5</td>
<td>46.03</td>
</tr>
<tr>
<td>ICRS</td>
<td>Adaptive stochastic method</td>
<td>183.579</td>
<td>16515</td>
<td>0.41</td>
</tr>
<tr>
<td>MCS</td>
<td>Multilevel Coordinate search</td>
<td>364.139</td>
<td>327697</td>
<td>39.06</td>
</tr>
<tr>
<td>GBSOLVE</td>
<td>Deterministic global optimizer</td>
<td>1179.464</td>
<td>649431</td>
<td>114.69</td>
</tr>
</tbody>
</table>

Table 5.1: Results of the global optimization methods

Hybrid approach and model identifiability

With the evidence that the optimization parameter can be performed best with Evolution strategies; Rodrigues-Fernandez [45] tried to find techniques to improve the fit and find statistics that can characterize the experimental data with respect to parameter identifiability. Although Evolutionary Strategies cannot guarantee global optimality, their robustness, plus the fact that in inverse problems they have a known lower bound for the cost function, makes them the best suitable candidates for the optimization. If we would use them for large-scale models it would take an unfeasible amount of time to get a good solution. To avoid this limitation Rodriguez came up with an idea is to combine global and local optimisation methods in order to reduce their weaknesses while enhancing their strengths. One would like to exploit the robustness of the global optimisation method with the rapid convergence of the local method. This idea can be considered as two sequential phases: first, an ES method is used to locate a solution in the vicinity of the global solution and, second, a fast local method is used to refine such solution until a pre-specified tolerance is met. For this global search phase an evolutionary strategy is a good option as it is robust, has good scaling properties and it has an almost self-tuning mechanism for the search parameters of the method.

The key issue in this scheme is to decide the amount of search to be performed by each method. This is done by defining stopping criteria, \( SC_1 \) and \( SC_2 \). The value \( SC_1 \) sets the switching point from the global to the local search and controls the robustness of the hybrid. Therefore, it must be chosen so as to ensure that the stochastic method will arrive at a point inside the radius of convergence of the deterministic method to the global optimum. The value \( SC_2 \) must be set such that the solution is close to the global optimum and controls the final computation time. To find these stopping criteria a simple heuristic rule was devised. Starting from the convergence curves obtained with some runs of the stochastic method, a number of
possible switching points are selected. From this data, one can find good $SC_1$ values which represent a compromise between robustness and efficiency.

A problem arising when estimating the parameters of the system from input-output data is often called the identification problem. There are two types of identifiability of the model. First the a priori structural identifiability problem: “Can we, under the ideal conditions of noise-free observations and error-free model structure and independently of the particular values of the parameters uniquely estimate the unknown parameters of the postulated model from the designed experiments?” This identifiability problem is hard to answer and in this study they use the posteriori or practical identifiability problem definition. This problem assumes that the model structure is exact, however, now data is sparse and noisy: “Can the unknown parameters of the postulated model be uniquely estimated from the (possible noisy and scarce) available data?” A problem with highly correlated parameters arises when a limited set of experimental, noise-corrupted data is used for parameter estimation. Under such conditions the uniqueness of parameter estimates predicted by the optimization, can no longer be guaranteed, because a change in one parameter can be compensated almost completely by a proportional shift in another parameter, still producing a satisfying fit between the experimental and model predictions. In addition, the optimization algorithm shows poor convergence when faced with this type of ill-conditioned optimization problem.

The question that we are trying to answer: “Does a small deviation in the found parameter set, have a considerable decrease of the fit as a consequence?” This question can be answered by the Fisher Information Matrix which quantifies the variance of the scoring/cost function [46]. The parameter estimation can be formulated as a minimization of the following quadratic cost function:

$$J(\tilde{\theta}) = \sum_{i=1}^{N} (\tilde{y}_i(\tilde{\theta}) - \tilde{y}_i)^T Q_i (\tilde{y}_i(\tilde{\theta}) - \tilde{y}_i) \quad (5.43)$$

where $\tilde{y}_i$ and $\tilde{y}_i(\tilde{\theta})$ are vectors of n measured model variables and model variable predictions at times $t_i$ ($i = 1$ to $N$) respectively. The matrix $Q_i$ is a square matrix of user-supplied weighting coefficients. The expected value of the cost function for a parameter set slightly different from the optimal one (global optimum) can be approximated by (Munack [47]):

$$E[J(\tilde{\theta} + \delta\tilde{\theta})] \approx \delta\tilde{\theta}^T \left[ \sum_{i=1}^{N} \left( \frac{\partial \tilde{y}_i}{\partial \tilde{\theta}} \bigg|_{t=t_i, \tilde{\theta} = \tilde{\theta}} \right)^T Q_i \left( \frac{\partial \tilde{y}_i}{\partial \tilde{\theta}} \bigg|_{t=t_i, \tilde{\theta} = \tilde{\theta}} \right) \right] \delta\tilde{\theta} + \sum_{i=1}^{N} \text{tr}(C_i Q_i) \quad (5.44)$$

where $C_i$ is the measurement error covariance matrix at time point $i$ and $\frac{\partial \tilde{y}_i}{\partial \tilde{\theta}}$ is the Jacobian of the model variables with respect to the parameters and are also called the sensitivities. The weight matrix $Q_i$ is typically chosen as $C_i^{-1}$ or the identity matrix. To optimize the practical identifiability we must maximize the difference between $J(\tilde{\theta} + \delta\tilde{\theta})$ and $J(\tilde{\theta})$. Let us assume that we find a parameter vector very close to the optimum and the differences of expected value of the cost function is small this most likely indicates that cost function is very flat around the optimum and that we will have trouble finding it by optimization methods. Hence, the practical identifiability of the model is low. To maximize this difference we should maximize the expression between the brackets, also called the Fisher Information Matrix (FIM).

$$F = \sum_{i=1}^{N} \left( \frac{\partial \tilde{y}_i}{\partial \tilde{\theta}} \bigg|_{t=t_i, \tilde{\theta} = \tilde{\theta}} \right)^T Q_i \left( \frac{\partial \tilde{y}_i}{\partial \tilde{\theta}} \bigg|_{t=t_i, \tilde{\theta} = \tilde{\theta}} \right) \quad (5.45)$$
This matrix is used to identify the lack of identifiability. Its inverse is the parameter estimation error covariance matrix of the best linear unbiased estimator:

\[ C = F^{-1} = \left( \sum_{i=1}^{N} \left( \frac{\partial \hat{y}}{\partial \hat{p}}_{t=t_i, p=p} \right)^T Q_i \left( \frac{\partial \hat{y}}{\partial \hat{p}}_{t=t_i, p=p} \right) \right)^{-1} \]  \hspace{1cm} (5.46)

The approximation of the objective function allows drawing lines of constant functional values in the parameter space. The FIM is symmetric; hence it has real eigenvalues and orthogonal eigenvector and can be used as a measure of the shape of the functional close to the optimal parameter estimates. We would like to design the experiments in such a way that the measurement data allows the unique or better identification of the parameters. The optimal experimental design problem is formulated as a dynamic optimisation problem, in which the objective is to find a set of input variables (controls) for the experiments that optimize the quality of the estimated parameters with respect to some metrics. In order to compare the efficacy of the experiment with respect to the parameter identifiability various functions of the FIM have been suggested as metrics.

<table>
<thead>
<tr>
<th>Name</th>
<th>Criterion</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>A-optimal design</td>
<td>Min(\text{trace}(F^{-1}))</td>
<td>Minimizing the trace of the inverse of the FIM matrix is equal to maximizing the length of random axes of the ellipsoid</td>
</tr>
<tr>
<td>Modified A-optimal</td>
<td>Max(\text{trace}(F))</td>
<td>Maximizing random eigenvalues of the ellipsoid results in maximizing the length of random axes of the ellipsoid</td>
</tr>
<tr>
<td>D-optimal design</td>
<td>Max(\text{det}(F))</td>
<td>Maximizes one or more random eigenvalues.</td>
</tr>
<tr>
<td>E-optimal design</td>
<td>\max_{\lambda_{\text{min}}} F</td>
<td>Increase the smallest eigenvalue, increase the smallest axis</td>
</tr>
<tr>
<td>Modified E-optimal</td>
<td>\min_{\lambda_{\text{min}}} \frac{\lambda_{\text{max}}(F)}{\lambda_{\text{min}}(F)}</td>
<td>Minimize the ratio of the largest and smallest eigenvalue. This results in a more spherical ellipsoid</td>
</tr>
</tbody>
</table>

The control variables in this case can be concentrations and temperatures, which can be varied during experiments to obtain information about the system. Mathematically the optimisation problem is:

\[
\text{Find } \bar{u}(t) \text{ that optimizes } J = \phi(F) \hspace{1cm} (5.47)
\]

\[
\text{Subject to } \dot{\hat{f}}[\dot{x}, \hat{x}, \hat{p}, \bar{u}, t] = 0, \dot{\hat{x}}(t_0) = x_0
\]

\[
\vec{h}[\hat{x}, \hat{p}, \hat{y}, \bar{u}, t] = 0
\]

\[
\vec{g}[\hat{x}, \hat{p}, \hat{y}, \bar{u}, t] \leq 0
\]

\[
\vec{x}^L \leq \hat{x} \leq \vec{x}^U
\]

\[
\bar{u}^L \leq \bar{u} \leq \bar{u}^U
\]

where the \( \phi \) is the criterion function, \( \hat{x} \) are the state variables, \( \bar{u} \) is the vector of input variables, \( \hat{y} \) is the vector of output variables or measurements, \( \dot{\hat{f}} \) is the system of ordinary-differential equality constraints with its initial conditions, \( \vec{h} \) and \( \vec{g} \) are the algebraic equality and inequality constraints. This non-linear problem is best solved by stochastic methods \([48]\). Finally we have a vector of control variables, which decorrelate the parameters and result in a cost function which can be optimized better.
Results

Exactly the same model is used as in the previous section to test this method on. Figure 5.21 illustrates the convergence curves which clearly shows that the performance as well as the computation time has improves with respect to the Evolutionary Strategies alone. The next step is to calculate the FIM from the sensitivities and the user-specified weight matrix. From the FIM we can calculate the parameter covariance matrix and the correlation matrix. The correlation matrix presented no correlation coefficients equal to +1 or -1 outside the diagonal indicating that all parameters are a posteriori identifiable. However, some pairs of parameters presented correlations close to +1 or -1, indicating the existence of identifiability difficulties.

![Figure 5.21: Convergence curves of the optimization. Black solid line is the optimization by SRES alone. Blue solid line is the global optimization method until the stopping criteria. Red dashed line is the local optimization.](image)

The left image of Figure 5.22 shows that the lack of correlation is reflected in a rather round contour plots of the cost function in the vicinity of the minimum, resulting in a unique parameter pair \((p_1, p_4)\). The right image of Figure 5.22 for the contour plots for the parameter pair \((p_1, p_6)\) with a high correlation. This contour plot has a long valley indicating that there are many combinations of \((p_1, p_6)\) giving the same value for the cost function. This problem of ill conditioning explains a part of the difficulties experienced by traditional local methods.

![Figure 5.22: Left: Cost function (surface plus contour lines) vs. parameters \(p_1\) and \(p_4\). Right: Cost function (surface plus contour lines) vs. parameters \(p_1\) and \(p_6\) indicating lack of identifiability.](image)

The parameters are estimated from 16 different time series generated from different combinations of pathway substrate and product concentration levels. These concentrations are taken as input variables of the system, such that we can optimize them to design more informative experiments. The criterion optimized in this study is the E-criterion and they constructed 16 and 10 new experiments to see what the improvements of the criterions are, shown in Figure 5.23.
The E-criterion increases significantly with the design of new experiments (as well as other criterions). This indicates that the parameters are decorrelated and that the identifiability difficulties have been decreased.

**Conclusion**

It is shown that the optimization problem can best be solved by global optimization methods. The Evolutionary Strategies performed best, but at the cost of large computation time. To solve this issue, Fernandez developed a hybrid optimization method capable of finding good solutions near the global optimum in significantly reduced time. We address the problem of parameter identifiability by calculating the FIM from the found parameter values and see the identifiability difficulties. With the use of optimal design criteria we can construct optimal experimental designs, which decrease these identifiability difficulties. The new estimate parameter set is more likely to improve with respect to accuracy and the optimization of the search space is less problematic.

**Limitation**

The optimization methods expressed here are limited by almost the same limitations defined for the optimization by the S-Systems representation. To not repeat myself I shall only state some additional limitations to these methods.

**Criteria:** One of the additional limitations is that identification of the stopping criteria, $S_C_1$ and $S_C_2$. These give the hybrid algorithm the robustness to find the global optimum, but they can be difficult to infer. One can never know if the stopping criterion $S_C_1$ is such that it finds the global optimum as it is not proofs that the found optimum is the global optimum.

**Prior knowledge of measurement noise:** Although noise can now be seen as a factor of identifiability difficulties, it is still not possible to integrate prior knowledge of the measurement noise in the optimization. It will try to fit the (noisy) experimental data without using prior knowledge.

**Scalability model:** The optimization procedures are moderately affected by the scalability of the model. If the number of model parameters increases the method is bound to take more computational time (even exponentially) and could be affected by the increase of local optima in the search space.
5.4 Filtering procedures

5.4.1 Introduction

The procedures explained in this section are purely time course methods. As already shown in previous chapters some problems are encountered in biochemical networks due to their non-linear nature. A very powerful approach to deal with these dynamic systems is the non-linear state space approach. To develop nonlinear state-space models for biochemical networks requires identifying the variables, their components and biochemical reactions which characterize the dynamics of the biochemical networks. In most cases, the variables in biochemical networks are concentrations of metabolites, enzymes and proteins or gene expression levels. These variables can yet again be divided into three groups:

- **Control variables**: Generally, the temporal external metabolite concentrations are taken as control variables, stored in the vector \( \mathbf{u}(t) \).
- **State variables**: The smallest variables that determine the state of the dynamic system and are often not easily accessible for measurement, hence they are called hidden, but essential for the evolution of the system, stored in the state vector \( \mathbf{x}(t) \).
- **Output variables**: These variables are the observed variables and are a function of the state variables and the parameters, stored in the variable vector \( \mathbf{y}(t) \).

Another influence on the system is the kinetic parameter. First we must give the state equations that define the dynamics of the biochemical network over time:

\[
\frac{d\mathbf{x}}{dt} = N \tilde{\mathbf{v}}(\mathbf{x}(t), \mathbf{u}(t), \mathbf{\theta}) \tag{5.48}
\]

The state equations shown in (5.48) assume an ideal situation. This assumption would in reality not be precise, thus is an approximation. This implies that the measurements taken shall differ slightly and this can be modeled by noise. To model this noise we state:

\[
d\mathbf{x} = N \tilde{\mathbf{v}}(\mathbf{x}(t), \mathbf{u}(t), \mathbf{\theta}) dt + G(t)d\mathbf{\beta}(t) \tag{5.49}
\]

where \( G(t) \) is the dispersion matrix and \( d\mathbf{\beta}(t) \) is the Brownian motion vector. Taking the time derivative gives:

\[
\frac{d\mathbf{x}}{dt} = N \tilde{\mathbf{v}}(\mathbf{x}(t), \mathbf{u}(t), \mathbf{\theta}) + G(t)\mathbf{\bar{w}}(t) \tag{5.50}
\]

where \( \mathbf{\bar{w}}(t) = \frac{d\mathbf{\beta}}{dt} \) is a Gaussian white noise process vector in the sense that at two different time points the vectors are uncorrelated. We have turned (5.48) into a stochastic ordinary differential equation by assuming that the dispersion matrix is the identity matrix. We define the observation equations where we assume that there is white Gaussian measurement noise, \( \tilde{\mathbf{v}}(t) \):

\[
\mathbf{y} = \mathbf{\tilde{h}}(\mathbf{x}(t), \mathbf{u}(t), \mathbf{\theta}) + \mathbf{\tilde{\nu}}(t) \tag{5.51}
\]
Because we cannot measure the variables in a continuous way we discretize (5.50-5.51) into difference equations. In a finite interval evolution (evolution from time point \( t_n \) to \( t_{n+1} \), which does not have to be constant) we state:

\[
\hat{F}(\hat{x}_n, \tilde{u}_n, \tilde{\theta}) = \hat{x}_n + \int_{t_n}^{t_{n+1}} N \tilde{u}(\tau) \hat{u}(\tau) d\tau \\
\hat{x}_{n+1} = \hat{F}(\hat{x}_n, \tilde{u}_n, \tilde{\theta}) + \tilde{w}_n \\
\hat{y}_n = \hat{h}(\hat{x}_n, \tilde{u}_n, \tilde{\theta}) + \tilde{v}_n
\] (5.52)

We denote the covariance matrix of the process noise \( \tilde{w}_n \) as \( Q_n \) and the covariance matrix of the measurement noise \( \tilde{v}_n \) as \( R_n \).

5.4.2 Linear Kalman filter

Introduction

In the case of linear systems, corrupted by white Gaussian noise, the Kalman filter is proven to be an optimal filter. The Kalman filter is a recursive algorithm that integrates all the provided information (system model (stoichiometric matrix, rate laws), measurement model, measurements, noise statistics, etc.) to estimate the current state of the system. The filter has two steps: prediction and update. The prediction step uses the system model and the information integrated so far to predict the a priori state variables. This step is also called the time update or evolution step as it propagates the current state variables and error. This estimate \( \hat{x}_k \) is different from the unknown true signal \( x_k \). To derive the estimate in an optimal manner the update stage uses the latest noisy measurements to correct the propagated state variables. This stage is also known as the measurement update, since it integrates the information brought by new measurements. While the Kalman filter is only optimal in the case of linear systems, several extensions to nonlinear systems exist: the Extended Kalman Filter (EKF) or the Unscented Kalman Filter (UKF).

Linear Kalman Filter

For the full derivation of the linear Kalman filter and the Bayesian interpretation we would like to refer to the supplement, section S5.4 and S5.5 respectively. This chapter gives a small description of the recursive Kalman Filter [49]. Consider the following linear model:

\[
x_k = F x_{k-1} + B u_{k-1} + w_{k-1} \quad (5.55) \\
y_k = H x_k + v_k \quad (5.56)
\]

With \( x_0 \), the initial state or mean of the state distribution, and \( P_0 \) the initial covariance. We have \( w_{k-1} \) (process noise) and \( v_k \) (measurement noise), Gaussian white noises with covariance matrix \( Q_k \) and \( R_k \) respectively. The goal is to recursively estimate (filter) the state \( x_k \) based upon the propagation of the current estimate \( \hat{x}_k \) and the update through available measurements.

**Prediction**: At time \( k \), the state \( x_{k-1} \) and its covariance \( P_{k-1} \) (which are known from the previous step) and the control input \( u_{k-1} \) are available. These are used to predict the next state and its covariance:

\[
\hat{x}_{k|k-1} = F x_{k-1} + B u_{k-1} \quad (5.57)
\]

\[
\hat{P}_{k|k-1} = F P_{k-1} F^T + Q_k \quad (5.58)
\]
**Update:** The predicted states are different from the unknown states and we wish to estimate it in an optimal manner. When the new measurements $y_k$ becomes available, it is used to compute the so-called Kalman gain matrix $K_k$ and correct the estimates.

\[
K_k = \hat{P}_{k|k-1}^T (H \hat{P}_{k|k-1}^T H^T + R_k)^{-1}
\]
\[
x_k = \hat{x}_{k|k-1} + K_k (y_k - H \hat{x}_{k|k-1})
\]
\[
P_k = (I - K_k H) \hat{P}_{k|k-1}
\]

In the end we have a Gaussian approximation of the posterior distribution, where the mean corresponds to the most likely set of state variables.

### 5.4.3 Extended Kalman Filter

**Introduction**

Several attempts have been made to extend/adapt Kalman filters for non-linear systems. One of the major issues with metabolic networks is that the rate laws are mostly non-linear, modeling the non-linear dynamics of the biochemical system. We may extend the use of Kalman filtering through a linearization procedure, resulting in the extended Kalman Filter (EKF). Such an extension is feasible by virtue of the fact that the Kalman filter is described in terms of difference equations in the case of discrete-time systems. The basic idea of the EKF is to linearize the state-space model at each time instants around the most recent state estimate. Once a linear model is obtained, the standard Kalman filter equations are applied. This section discusses the EKF as a state-of-the-art methodology in parameters estimation in metabolomics [50]. Although this kind of filter was already used in other fields of control system identification it has not frequently been used in metabolic modeling, due to some disadvantages (discussed later).

**Methods**

We use the same state-space model as in (5.53-5.54), but with a small extension. What we would like to do is use the EKF for dual estimation, where we want the estimate the state variables as well as the parameters from input variables and noisy observations. We model this by taking the parameters as additional states and augment them to the state variables. The parameter vector corresponds to a stationary process, with identity state transition matrix, by uncorrelated white Gaussian noise, $\mu_k$, with covariance matrix $\Phi_k$. Setting this matrix has a large influence on the convergence of the algorithm to the true parameters. We define the new state vector as:

\[
\hat{z}_{k+1} = \begin{bmatrix} \hat{x}_{k+1} \\ \hat{\theta}_{k+1} \end{bmatrix} = \begin{bmatrix} f(x_k, u_k, \theta_k) \\ \theta_k \end{bmatrix} + \begin{bmatrix} w_k \\ \mu_k \end{bmatrix} = g(z_k, u_k) + \xi_k
\]

The measurement vector is represented mathematically as:

\[
y_{k+1} = h(z_k, u_k) + v_k
\]
We define \( \hat{z}_{k|k} = \begin{bmatrix} \hat{x}_{k|k} \\ \hat{\theta}_{k|k} \end{bmatrix} \) as the new estimated state vectors and define the linearized transition (system) model and measurement model as:

\[
F_k = \begin{bmatrix} \frac{\partial f}{\partial x} & \frac{\partial f}{\partial \theta} \\ 0 & I \end{bmatrix}_{x=\hat{x}_{k|k}, \theta=\hat{\theta}_{k|k}} \quad \text{and} \quad H_{k+1} = \begin{bmatrix} \frac{\partial h}{\partial x} & \frac{\partial h}{\partial \theta} \\ \end{bmatrix}_{x=\hat{x}_{k+1|k}, \theta=\hat{\theta}_{k+1|k}, u=u_{k+1}} (5.64)
\]

**Prediction:**

\[
\hat{z}_{k|k-1} = g(\hat{z}_{k|k}, u_k) \quad (5.65)
\]

\[
P_{k+1|k} = F_k P_k|k F_k^T + \Psi_k \quad (5.66)
\]

\[
\Psi_k = \begin{bmatrix} Q_k & 0 \\ 0 & \Phi_k \end{bmatrix} \quad (5.67)
\]

**Update:**

\[
K_{k+1} = P_{k+1|k} H_{k+1}^T (H_{k+1} P_{k+1|k} H_{k+1}^T + R_{k+1})^{-1} \quad (5.68)
\]

\[
\hat{z}_{k+1|k+1} = \hat{z}_{k+1|k} + K_{k+1}(\bar{y}_{k+1} - \tilde{h}(\hat{z}_{k+1|k}, \bar{u}_{k+1})) \quad (5.69)
\]

\[
P_{k+1|k+1} = (I - K_{k+1} H_{k+1}) P_{k+1|k} \quad (5.70)
\]

**Results**

The EKF was tested on the JAK-STAT pathway as. For each time point a large number of cells were taken from a pool of BaF3C cells. The state variables include concentrations of unphosphorylated STAT5 \((x_1)\), tyrosine phosphorylated monomeric STAT5 \((x_2)\), tyrosine phosphorylated dimeric STAT5 \((x_3)\) and nuclear STAT5 \((x_4)\). Only concentrations of tyrosine phosphorylated STAT5 in the cytoplasm and total STAT5 were measured for 16 time points and the measurements of Epo-induced tyrosine phosphorylated EpoR were used as control variable. The left image of Figure 5.24 illustrates the prediction and the observed concentration of the total concentration of STAT5, it shows the predicted values lie close to the true values. The right image of Figure 5.24 illustrates concentration of the state variables. We cannot state if the predicted curves follow the real curves as they are not observed. It was shown that the estimated parameters in this test was very close to estimated values by other studies [50].

![Figure 5.24](image.png)
Conclusion:

To estimate the parameters of a non-linear state-space model we extended the Kalman filter by linearizing the state-space model. Using the observed data, to update the estimated values optimally, we can recover accurately the parameters of the model. One disadvantage is that for a fast convergence one needs to initialize the state variables (including parameters) close to their real values. If this is not the case we need a lot of data to estimate the true parameters and state variables. The limitations of the Extended Kalman Filter will be discussed in the limitation sections.

5.4.4 Unscented Kalman Filter

Introduction

Using the EKF, the state distribution is approximated by a joint Gaussian distribution, which is then propagated analytically through the first-order linearization of the nonlinear state-space model. This can introduce large errors in true posterior mean and covariance of the transformed Gaussian distribution, which may lead to suboptimal performance and sometimes divergence of the filter. To avoid this problem a variation of the Kalman filter was designed, called the Unscented Kalman Filter (UKF). The state distribution is again approximated by a joint Gaussian distribution but is now represented using a minimal set of carefully chosen sample point, called sigma points. These points completely capture the true mean and covariance of the state distribution when propagated through the non-linear system. It is accurate to a second order (Taylor series expansion) approximation for ANY non-linearity. If the state distribution is assumed to be Gaussian it can be even approximated by a third-order approximation. The methodology of the UKF is more difficult than that of the EKF and shall be omitted from this chapter, but if the reader would like to know the mechanism one can look in the supplement section S5.6.

Results:

A test of the UKF for system identification was performed on the repressilator model, shown in Figure 5.25. The initial concentrations of the components were drawn independently from a uniform distribution on [0,100] and the total time interval was taken to be 20 seconds. Observations noises are added to three observed variables to mimic gene expression data and were taken equal to 20% of the standard deviation of the state variables. The measurement interval is 0.2 seconds, which is really fast and unlikely to be producible in real-life, resulting in 100 time points. The mRNA concentrations are known, but the protein concentrations are hidden. We would like to learn the following six parameters: $v_1^{\text{max}}, v_2^{\text{max}}, v_3^{\text{max}}, k_{12}, k_{23}, k_{31}$. In order to estimate the true parameters a multi-start approach was used by sampling 50 different initial states for the parameters and state variables from the prior state distribution $\pi(\theta, x_0)$. From these multiple runs the final mean values and covariances matrices are averaged.

Figure 5.25: Repressilator together with its system of differential equations.
Figure 5.26 illustrates the estimations of the six parameter values. Although the initial parameters are not near their true values, we can still estimate the true parameters with a good accuracy after a large number of experiments. Note that the time interval in this figure is longer than the predefined 20 seconds.

A limitation is that the UKF needs a small time interval to accurately estimate the state variable, due the fact that the Kalman Filter is an ‘on-line’ method. The mean-squared error between the predicted concentrations and the observed concentrations of the hidden variables increase significantly as illustrated in Figure 5.27.

5.4.5 Limitations

In the case of a linear process model corrupted by zero-mean Gaussian noise, the Kalman filter is the optimal estimator. This is because of the property that a Gaussian process going through a linear system will remain Gaussian. However, propagating through a nonlinear system this Gaussian distribution will not likely remain Gaussian.

**Noise quantification:** It is difficult to quantify the process noise and measurement noise. In both studies it was assumed to be known but this is not realistic. The noise covariance matrices can vary over time, but are in all studies assumed to remain constant. The only way to quantify this is making assumptions about the type and extent of the noise.

**Noise identification:** Some measurement noise does not have to be white noise and can be considered colored noise. This type of noise cannot be processed by Kalman Filters.

**Independence assumption:** A simplifying assumption is that the process noise and measurement noise are uncorrelated and in practice this might not be the case. This implies that the Kalman Filter models incorrect phenomenon and could result in suboptimal estimations.
**Local optima:** A fact is that the Kalman Filter can get stuck in local optima. We can try to avoid this problem using multiple starts, taking different initial settings from the prior state distribution. A problem arising is that generally this state distribution is not known and must be estimated, which heavily influences the convergence of the algorithm to the true parameters. Slow convergence means many time point measurements are needed, which is in practice is hard.

**Measurement interval:** Both papers simulated pseudo-experimental data from the differential equations with a small measurement interval of 0.2 seconds. This is very unlikely as current state-of-the-art methods take measurements in the order of seconds. The Kalman Filters thrive on the fact that the measurement interval is small and thus generate a lot of data points. Normally one has, if lucky, somewhat between 10 to 20 measurements, and the estimation with the Kalman Filter is likely to be poor. The Kalman Filters are normally used as a ‘real-time’ processor, implying that the time interval between the measurements is very small. This is certainly not the case in metabolic engineering, resulting in suboptimal estimates.

**Scalability:** The most important limitation of the Kalman filters is the scalability. Due to the experimental cost of measuring kinetic data, the number of time points and replications are limited, which can affect the size of the networks that the Kalman Filters can fit accurately. There is also a relation between the number of measurable model variables and the number hidden state variables/parameters which we can estimate with the Kalman filter. Both studies had a small number of parameters and we had a few measurable output variables. These parameters can easily be inferred, but as we go to larger scale networks we are very limited. The limitation might be such that these methodologies cannot be used for large-scale networks.
Discussion, future work and open problems

6.1 Limitations

In this report we have divided parameter estimations methods for metabolic networks into three classes: (I) steady-state methods (parameter space restriction methods), (II) time course/steady-state methods (optimization), (III) time course methods (filtering). Each of these methods has their own limitations and advantages. In this section we will discuss the major limitations for each class of methods. The limitations given are based on the assumption that we would like to learn more about the biochemical system (system identification).

<table>
<thead>
<tr>
<th>Limitation</th>
<th>Parameter space restriction</th>
<th>Optimization</th>
<th>Filtering</th>
</tr>
</thead>
<tbody>
<tr>
<td>Integration prior knowledge of measurement noise</td>
<td>+</td>
<td>-</td>
<td>+</td>
</tr>
<tr>
<td>Network scalability</td>
<td>+</td>
<td>±</td>
<td>-</td>
</tr>
<tr>
<td>Biological interpretability</td>
<td>+</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Optimal Experimental Design</td>
<td>-</td>
<td>+</td>
<td>-</td>
</tr>
<tr>
<td>Amount of data needed for good estimation</td>
<td>+</td>
<td>±</td>
<td>±</td>
</tr>
<tr>
<td>Complexity of model it can handle</td>
<td>-</td>
<td>±</td>
<td>±</td>
</tr>
</tbody>
</table>

+: Not limited  
±: Limited to some degree  
-: Limited

Integration prior knowledge of measurement noise:

It is very likely that the measurements are corrupted by noise. Measurements in metabolic engineering take much time and effort, resulting sometimes in noise of 30% of the true value. It would thus be desirable to integrate prior knowledge of the measurement noise into the estimation procedure. The parameter restriction methods and Kalman Filters integrate this prior knowledge, although under the assumption that it is normally distributed (which has not been proven). On the other side of the spectrum we have the optimization procedures which do not integrate this prior knowledge at all. They will try to fit the (noisy) experimental data without using this knowledge resulting most likely in sub-optimal estimates.

Network scalability:

The parameter space restriction methods can easily handle large-scale networks and are not rendered unfeasible if the size of the network increases. The optimization procedures are limited by the scale of the network as the dimensionality of the parameter space increases, resulting in more local optima and larger computation time to find the global optimum. A method developed by Kuepfer [52] uses a combination of optimization procedures with parameter restriction methods and is less limited in network scale. The filtering methods are very limited in the size of the networks they can handle. This size is expressed by the relation between the observed variables and the hidden states (including the parameters).
Biological interpretability:

The parameter space restriction methods are bottom-up procedures where different pieces of evidence are put together to estimate the parameters. The parameters cannot be estimated perfectly from this evidence, so these methods quantify the uncertainty with distributions or restrictions. From these distributions we can calculate all different kinds of properties of the system, such as parameter sensitivity or system behavior, and gives us the possibility to interpret the estimates biologically. They do not provide one single “best” parameter set. This is not the case for top-down approaches, such as optimization and the Kalman filters, which give one single parameter set and does not give us an opportunity to gain knowledge about the system. These kinds of methods must be used if one just wants to control the system.

Optimal Experimental Design:

We can perform optimal experimental design to improve the practical identifiability. The optimization methods were the only class of methods where the optimal experimental design could be applied due to the cost function, giving it in this field an advantage over the other two classes of methods.

Amount of data needed for good estimation:

With the parameter space restriction methods, given a small number of steady state data originating from different conditions, we can already restrict the parameter space or calculate narrow joint distributions very accurately. The top-down approaches will only find good estimates if the initial parameter values are already close to the true values or secondly if we have a large number of time points. The acquisition of this data can cost a lot of time, effort and even money.

Complexity of the model it can handle:

One limitation of the parameter space restriction methods is that they are based on simplifying assumption of the rate laws. Although most enzymatic reactions follow a similar mechanism, other enzymes can display more complex behavior. The optimization and Kalman filtering procedures do not make simplifications and can model the non-linear state equations. A problem is that this is only to a certain degree; for the optimization procedures a more complex model leads to complex search spaces and Kalman filters can only estimate non-linear systems to second order. Also a complex model generally implies more parameters, leading to scalability limitations.

6.2 Future work and open problems

In this section I will give some directions for future work and discuss some problems which have not yet been solved.

Network decomposition:

One of the limitations of most methods is the size of the network they can fit. Some methods cannot fit large networks due to computational cost, whereas others suffer from the enormous amount of parameters to be estimated and data they need. Automatic network decomposition could help here. An advantage in the modeling of metabolic networks is that we know its complete structure, i.e. its stoichiometric matrix. We can decompose the network into smaller sub-networks, as each sub-network has less parameters to estimate. Another advantage is that we could estimate the parameters for each sub-network with methods to the researchers liking. Linear sub-networks can be modeled by rather simple techniques, while other non-linear sub-networks can be modeled by more complex methods such as Kalman Filter. There are
automatic methods to decompose metabolic networks in the form of Petrinets [53]. The problem with these is that they cannot decompose networks if allosteric regulation is incorporated in the model. Figure 6.1 illustrates this problem by a small metabolic network which we would like to decompose. The decomposition according to the dashed line is undesirable, because it cuts off the allosteric regulations influencing the reaction rate. The decomposition of the network such that it takes into account all interactions has not yet been found.

![Figure 6.1: Decomposition of metabolic network. Nodes are metabolites, edges are reactions, blunt interaction is inhibition, and arrowed interaction is activation.](image)

**Particle filters**

A method proposed in multiple papers is the particle filter. This is an old concept in system control and system identification, but has just been proposed as a method to find metabolic parameters accurately. Kalman filters have the disadvantage that the distribution of the random variables is assumed to be Gaussian, which is not always the case. Furthermore, the result of propagating a Gaussian through a nonlinear function is in most cases not a Gaussian anymore. To cope with this problem, particle filters represent the posterior distribution by weighted samples, rather than computing it analytically. Although this method is too complex to discuss here, we will give a small indication how it works. From the true posterior distribution we can randomly pick a large number of samples or particles that can represent this distribution, because the true distribution cannot be expressed by its Bayesian formula. The Bayesian Framework is transformed such that the particles are drawn from a suitable proposal distribution. Particles are then propagated such that they represent the propagated posterior distribution. The particle filter has some advantages over the Kalman filter as it can handle any type of nonlinear model and any distribution [49]. They also have some disadvantages:

**Computational time:** To reconstruct the posterior distribution we need a lot of samples (even infinite if we want to have a perfect reconstruction). With today’s computational powers and the possibility of concurrent programming we can model networks with 20-30 parameters.

**Distribution interpretability:** A problem is that the state distribution may not be a Gaussian, so taking the mean as most likely parameter vector might be false due to multimodality.

**Proposal distribution:** The proposal distribution from which the samples are drawn is unknown in advance. This implies that the researcher must specify a suitable distribution with assumptions. Specifying a correct proposal distribution results in a lower number of needed particles, thus reducing the computation time.

Particle filters need large computation time, which increases exponentially with network size. The decomposition of the network is a possibility for improvement for this method. The choice of proposal distribution is currently still an active field of research, just as the schemes for resampling. This indicates that there are still open problems in this field.
Currently there are no papers in which particle filters are tested on metabolic networks, but they are seen as tool for the future in many papers. A not (yet) published study has been performed by Quanch [54], where he used the Particle Filter on the repressilator. It showed that the particle filter outperformed the Unscented Kalman filter, but it needed significantly more time.

**B-spline estimators**

This type of estimator fits the time data by a non-linear regression. The generalization of the fit can be controlled by some parameters (in this case the number of knots). By generalizing the observation data to a fit or curve we can estimate parameters from this curve [55]. This is a rather new method used for finding parameters of ODE’s. It was not invented specifically for biochemical systems, but has been tested on these systems with positive results. The development of new types of estimators shall in the future be studied.

**Model reduction**

It would be desirable to reduce the model to fewer reactions or parameters. Some rate laws such as lin-log kinetics [28] need fewer parameters to be estimated while they can approximate the enzyme mechanism accurately. The development of approximative methods to reduce the model to fewer reactions or parameters is still a field that needs to be developed further.

**Smoothers**

Suppose that we are given a set of data over the time interval $0 < k \leq N$. Smoothing is a non-real-time operation in that it involves estimation of the state vector $\tilde{x}_k$ for $0 < k \leq N$, using all available data, past as well as future [56]. To determine the optimum state estimates $\tilde{x}_k$ for $0 < k \leq N$, we need to account for past data $\tilde{y}_j$ defined by $0 < j \leq k$ and future data $\tilde{y}_w$ defined by $j < k \leq N$. The estimation pertaining to the past data, forward filtering, can be performed by a Kalman filter. To deal with the issue of state estimation pertaining to the future data, we used backward filtering, which starts at the final time $N$ and runs backwards. Finally a smoother is used to combine the two estimate state estimates. It has been shown that these smoothing procedures give generally better results than Kalman filters and do not need to compute the state estimates in real-time [57].
6.3 Proposals

This section gives some proposals that have come up during the literature survey. They are based on a possible synergy of methods.

**Optimization:**

For the optimization methods the convergence and performance with respect to the error is very dependent on the initialization of the parameter vector. Normally these were sampled from a uniform or normal distribution constructed using knowledge from the true parameters values (they were always in the range of initialization). Furthermore the range of initialization was kept tight such that the randomly drawn parameters were already close to the true parameters. I would like to propose to use the methods of Liebermeister [29] to create a more reasonable initial distribution. The joint distribution, estimated from the steady-state data, tells use the feasible areas for the true parameters and thus can be used to initialize from which in turn could result in a fast convergence to the global optimum.

**Kalman filter:**

The Kalman Filter needs a joint Gaussian distribution for the initial state to start the algorithm. As shown a multi-start approach is taken to avoid local optima. In the literature these joint distribution as mostly just constructed without any sound evidence of the true parameters and their uncertainty. Also here, the joint Gaussian distribution generated from Liebermeister [29] could play a significant role in proper initialization. In this case this would lead to a fast convergence and the need of less time points and effort.

**Particle filter:**

Just like the Kalman filter the particle filter needs an initial probability distribution for the initial parameters. The same proposal as with the Kalman filter applies here, but here we can go further. One of the more difficult tasks when using particle filters is the choice or identification of the proposal distribution. We could use the Gaussian posterior distribution from Liebermeister as a proposal distribution. This distribution is “updated” by data and will have a similar shape and support as the real posterior distribution. This choice could lead to need of fewer particles during the evolution of the system, hence reducing the computational cost.
6.4 Discussion

In this chapter we have defined the major limitations of the methods and gave directions for future work. We conclude with a brief discussion about the classes of methods and indicate which of them should be further investigated.

The choice of method is very dependent of the question if one would like to control the system or learn more about its behavior. The parameter space restriction methods generally give more information about the behavior of the system. It does not give one single best parameter set, but quantifies the uncertainty of the estimation by a posterior Gaussian distribution. As shown we can also calculate different kinds of properties of the biochemical model based on these distributions. The optimization and Kalman filter procedures will only give one parameter set which reproduces the time-course data as closely as possible. These parameters are hard to interpret, as these methods generally are used only to control the system and not gain any knowledge. Under the assumption that we would like to gain knowledge about the model, we advise to further investigate the parameter restriction methods.

One of the major limitations of these methods is that they all make simplifying assumptions about the enzyme kinetics. This could lead to sub-optimal estimates as some enzymatic reactions have a complex mechanism. To solve this problem we can make use optimization procedures and Kalman filters as they can estimate parameters for non-linear systems. Both procedures are limited in the scale of the network they can handle, implying that they can only fit a model of a certain size. Furthermore the optimization procedures are handicapped by complex search spaces for large non-linear models. The Kalman filters can estimate parameters for high non-linear models but make some assumptions about the state distribution and is limited to the number of parameters it can estimate from the experimental data. To avoid this problem the use of particle filters is suggested by many studies, which are related to the Kalman Filter. A downside of the particle filter is the computation time, but with faster computers and concurrent programming we can solve this problem. Furthermore decomposition of the metabolic network can improve this method. Also smoothers are suggested as they do not need real-time data and generally give better estimates of the parameters. We have seen a surge in papers using smoothers and filters to perform system identification. With the increase of computational power and the possibility of more accurate and faster (smaller measurement interval) measurements these methods can generally give better estimations. Also a large number of new filtering and smoothing methods have been developed in the classical field of control system theory, ready to be used for parameter estimation in biochemical networks. This evidence shows that an investigation in this area should result in sophisticated (possibly new) methods which can give good estimates. Therefore we advise to put effort in investigating smoothers and filters.
6.5 Conclusion

The field of metabolic engineering has come a long way. Driven by the need for understanding the behavior of metabolic pathways and networks, dynamic modeling has been extended intensively. The large number of tools these days gives researchers the capability to investigate the dynamical behavior of the biochemical systems with a large number of industrial applications.

One of the downsides of these dynamic models is that they are governed by complex mechanisms, which are diverse and not always fully understood. To model these mechanisms one has to construct approximative mathematical models. The dynamics of the model are governed by a large number of parameters, such as reaction rate constants. One of the major problems in the field of metabolic networks is the identification of these parameters, called system identification. Actually, this field is already quite old. Although many methods have already been developed, the need of new or improved methods for parameter estimation is ever present.

This literature survey was performed to research state-of-the-arts methods currently used. With the identification of the most important rate laws modeling enzymatic mechanisms, came the explanation of the different types of parameter estimation methods. These methods were divided into three classes: (I) steady-state methods (parameter space restriction methods), (II) time course/steady-state methods (optimization), (III) time course methods (filtering). The first class estimates parameters based on steady-state data and rather than giving one parameter vector it will give a distribution of likely parameters that describe the given data. From these distributions we can gain knowledge about the system and calculate properties of the model.

The second class of methods tries to optimize an objective function by tweaking the parameter vector. This approach results in a parameter vector that reproduces the observation as good as possible. A limitation for these methods is that the problem is highly multimodal and takes a lot of computation time. An advantage is that one can design optimal experiments such that one has more discriminative power over the parameters.

Based on nonlinear state-space models, filtering methods can estimate the parameters and even hidden variables, which are mostly metabolite concentrations not measurable, accurately as long as the initial parameters are close to the true parameters and there is a lot of time series data. An advantage of this method is that it can model complex systems, but is yet limited in size of the network it can fit.

Some future directions have indicated that research in this field is far from done. Decompositional and approximative methods could make methods earlier thought of as unfeasible, such as the particle filter, usable. With increasing computational power and the capability to perform rapid sampling the use of previously non-feasible methods has become possible. Many (new) methods developed for control systems are making their way into systems biology, indicating that the communities have for too long communicated poorly.
References


