METABOLIC NETWORK MODEL IDENTIFICATION — PARAMETER ESTIMATION AND ENSEMBLE MODELING

JIA GENGJIE

(B. Sci. University of Science and Technology of China)

A THESIS SUBMITTED

FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

IN CHEMICAL AND PHARMACEUTICAL ENGINEERING (CPE)

SINGAPORE-MIT ALLIANCE

NATIONAL UNIVERSITY OF SINGAPORE

2012

DECLARATION

I hereby declare that this thesis is my original work and it has been written by me in its entirety. I have duly acknowledged all the sources of information which have been used in the thesis.

This thesis has also not been submitted for any degree in any university previously.

Jia Gengjie

JIA GENGJIE

2 August 2012

ACKNOWLEDGEMENTS

In my case, truth pursuit in the research has always been a process of path finding and problem solving one after another, which has trained me with creative ideas, critical thinking, analytical mindset and computational skills. In this regard, any of my achievements would have been impossible without the supports I received on the way.

First of all, my words fail to express my sincere gratitude to my supervisors in ETH, MIT and NUS. Dr. Rudiyanto Gunawan, who has brought me to the field of Computational Systems Biology, is always a patient teacher and a kind friend to me. He creates many opportunities for his students to attend overseas studies, conferences and seminars. His trust, encouragement and guidance are the great boost to my studies, and I have learnt so much from him by discussing the issues in my research and career planning. I am also great thankful of the inspirable suggestions and guidance from Prof. Gregory N. Stephanopoulos, who has led me to the field of Metabolic Engineering. I would like to express my great gratitude to Dr. Mark Saeys for constantly sharing his invaluable experiences with me on research and technical trainings.

I appreciate the guidance from A/P Heng-Phon Too, from whom I have learnt innumerable insights on research through collaborations and discussions with him. I also thank Dr. Saif A. Khan and Prof. Patrick S. Doyle for serving in my thesis examination committee and advising for my research work.

In addition, I shall thank all my friends, especially my lab mates: Suresh Kumar Poovathingal, Thanneer Malai Perumal, Sridharan Srinath, Lakshminarayanan Lakshmanan, Zhi Yang Tam, Yang Liu and S. M. Minhaz Ud-Dean, who have been such great companions during my postgraduate studies and encourage me to improve every day. Thanks for creating such a wonderful working environment in the lab, and I have benefited so much from the discussions with them during group meetings, even lunch and dinner time.

I would like to acknowledge the funding supports from Singapore-MIT Alliance (SMA) and ETH, and to thank Ms. Juliana Chai and Ms. Lyn Chua for their unrelenting technical and administrative help. I also appreciate the department of Chemical and Biomolecular Engineering (ChBE), NUS, for offering me necessary facilities and research seminars. My gratitude should also be given to my teachers: A/P. Lakshminarayanan Samavedham, A/P. Kai Chee Loh, Dr. Chitra Varaprasad, Prof. Raj Rajagopalan and so forth.

As for my publications, I appreciate the help from Prof. Eberhard O. Voit for sharing model formulations and measurement data for case studies, and thank Dr. Jose A. Egea and Prof. Julio R. Banga for their assistance in using SSm GO toolbox.

Last but most importantly, I thank my parents and wife for strong and constant supports, promoting my growth in the past, present and future.

.

TABLE OF CONTENTS

SUMMARY

Metabolic Engineering employs targeted alterations of metabolism in microbial organisms for biochemical production. In practice, the re-engineering of cellular metabolism involves a cyclic procedure, including strain construction, strain characterization, metabolic systems analysis and strain design. Mathematical modeling plays an important role in this procedure, in describing system dynamics and predicting system responses upon perturbations. Here, kinetic models are especially useful when the system dynamics and regulatory are of particular interest in the study.

Recent advances in molecular biology techniques have permitted the simultaneous collection of large quantities of metabolic network information, such as time-course measurements of gene expression, protein abundances and metabolite concentrations. The underlying information about the metabolic network in those data, however, is implicit and requires subsequent extraction, which can be facilitated by building mathematical models. Constructing kinetic models from time-series data is challenging and parameter estimation remains a bottlenecking step in this process. The challenges can be categorized into four areas: data-related, model-related, computational and mathematical issues. To tackle these issues, extensive efforts have previously been made in developing various support algorithms as well as optimization methods. Nevertheless, numerous problems still remain unsolved, constituting significant research gaps in the field.

Motivated by some of the issues in the kinetic metabolic modeling, the present PhD project focuses on the development of efficient model identification methods and framework to capture model uncertainty. More specifically, the methods are developed to address three common issues related to the estimation of parameters in kinetic metabolic models, namely (1) missing information of some metabolites, (2) high computational demand associated with stiff ordinary differential equations (ODEs) and large parameter search space, and (3) degrees of freedom in the model due to larger number of metabolic fluxes than metabolites. These problems often led to challenging parameter estimations for which existing algorithms either fail or become impractical due to high computational requirement. In this thesis, I present three computationally efficient algorithms for the purposes of (1) estimating parameters from incomplete metabolic profiles using a two-phase dynamic decoupling method, (2) estimating parameters using an incremental approach, and (3) constructing a kinetic model ensemble using an incremental approach. The efficacy of the three proposed methods has been demonstrated through applications to a few case studies (artificial and real metabolic pathways) and through comparisons with existing methods.

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1 : INTRODUCTION

1.1 Problem Formulation

1.1.1 Metabolic Engineering and Mathematical Modeling

Chemical industry is undergoing a dramatic change motivated by an increasing demand for sustainable processes for the production of fuels, materials and pharmaceuticals. As traditional synthetic routes often face numerous problems due to increasing raw material costs, environmental constraints and sustainability requirements, biotechnology, in conjunction with genetic engineering, offers a sustainable and environmental-friendly solution [1]. With the invention of recombinant DNA technology, microbes like *Escherichia coli* and *Saccharomyces cerevisieae* (yeast) can be used to produce valuable products through modification or introduction of some biochemical reactions. This is the essence of Metabolic Engineering [2], an area that has garnered global attention from academia to industry and has experienced unprecedented growth in the last fifteen years. Within this frame, many metabolites with great therapeutic and economic values have been produced, such as Lycopene [3], Artemisinin precursors [4], Benzylisoquinoline alkaloids [5], L-valine [6] and Isoprenoids [7].

Metabolic Engineering relies on the knowledge of cellular metabolism and its regulation, and the technology encompasses two defining steps: analysis and synthesis, relying on an integrated view of metabolic pathways instead of

individual reactions [8]. Consequently, mathematical modeling of metabolic networks has played an important role in predicting and analyzing microbial metabolism *in silico*, from which metabolic manipulations can be rationally designed and screened prior to actual experiments. The value of mathematical models has been clearly shown in understanding essential qualitative features of biological systems, capturing essential quantitative characteristics of experimental data, describing interactions within complex systems, correcting conventional knowledge, and predicting possible system responses upon different perturbations, all of which have been widely documented in prior studies [9].

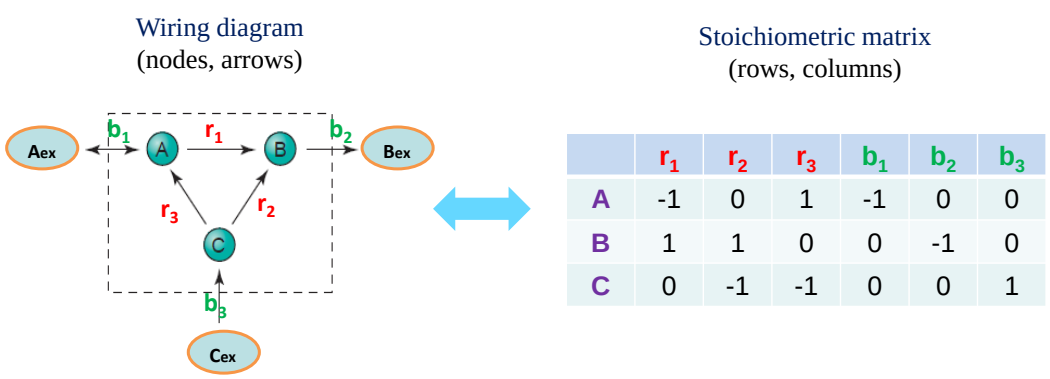
Mathematical models of metabolic pathways are typically constructed based on mass balances of intracellular metabolites, written as a set of ordinary differential equations (ODEs) as follow:

$$\dot{\mathbf{X}} = \mathbf{S}\mathbf{v}, \tag{1.1}$$

where $\mathbf{X} = \{X_1, X_2, ..., X_m\}d$ is the vector of the concentrations of m metabolites, $\mathbf{v} = \{v_1, v_2, ..., v_n\}$ is the metabolic flux vector, and $\mathbf{S}$ denotes $m \times n$ stoichiometric matrix [8,10]. In general, metabolic fluxes depend on both metabolite concentrations $\mathbf{X}$ and (unknown) kinetic parameters $\mathbf{p}$, i.e., $v_i = v_i\ (\mathbf{X}, \mathbf{p})$. Such kinetic ODE models can be used directly in analysis, or by assuming steady state, simplified to an algebraic stoichiometric model $\mathbf{S}\mathbf{v}=0$. Below, I will discuss these two models in greater detail.

1.1.2 Stoichiometric Models

The stoichiometry of metabolic pathways describes the topology of metabolic networks, which can be visualized by a wiring diagram of metabolic pathways. Conventionally, metabolites are represented by nodes and metabolic fluxes by directed edges or arrows. Vice versa, given a topological wiring diagram with *m* metabolites and *n* fluxes, a stoichiometric matrix can be constructed, in which the rows correspond to metabolites and the columns to reactions that affect the said metabolite concentration (see Figure 1.1). That is, $\mathbf{S}_{ij}$ is the stoichiometric coefficient of the *i*-th metabolite participating in the *j*-th reaction. The construction of this matrix constitutes one step in model identification that translates the biological network diagram into mathematical terms [11].



	r_1	r_2	r_3	b_1	b_2	b_3
A	-1	0	1	-1	0	0
B	1	1	0	0	-1	0
C	0	-1	-1	0	0	1

Figure 1.1. A wiring diagram and stoichiometric matrix of a metabolic network.

Under steady-state assumption, giving **Sv**=0, several methodologies have been developed to exploit mathematical descriptions for cell metabolism, which are based on different assumptions (e.g., maximal growth rate, maximal productivity

or minimal nutrient consumption), have different purposes (e.g., to analyze a network or to make predictions upon perturbations), and adopt different mathematical frameworks (e.g., linear algebra or convex basis). Basically, these methods can be classified into two branches: those for determining feasible flux solutions (e.g., Metabolic Flux Analysis and Flux Balance Analysis) and those focused on the properties of the entire space of possible flux distributions (e.g., Extreme Pathway Analysis and Elementary Mode Analysis) [12,13] (see Figure 1.2).

Metabolic Flux Analysis (MFA) has been commonly used to predict the intracellular fluxes, based on a set of measured extracellular fluxes from which the information is sufficient enough to reduce the solution space of the system to finitely many points [14,15]. Mathematically speaking, this requires a determined system, of which its linearly independent constraints are sufficient to uniquely identify the unmeasured fluxes. For an underdetermined system, Flux Balance Analysis (FBA) can be applied to predict flux distributions. As there are more fluxes than metabolites in a typical metabolic pathway, there exist an infinite number of solutions to the steady-state model $\mathbf{Sv}=0$. To select the most biologically relevant flux distribution among the set of feasible solutions, the FBA relies on the assumption that cells have evolved to achieve an optimal status owing to evolutionary pressure [15,16]. For instance, the most common hypothesis in FBA is that microbes regulate their metabolism to maximize the growth of themselves [17,18]. The advantage of FBA is that only the stoichiometric matrix information is needed to predict the metabolic fluxes.

Nevertheless, these flux predictions greatly depend on the optimality assumption, which may not stand in the same organisms all the time and even after a genetic modification. Furthermore, it is not clear whether the same optimality condition can be maintained by different organisms.

Other analyses based on the steady-state assumption have also been formulated, including Extreme Pathway Analysis (EPA) and Elementary Mode Analysis (EMA) [19]. Built on the concept of convex analysis, in these analyses, one compute the basis flux vectors, called extreme pathways [20,21] or elementary modes [22,23], from which all the solutions of $\mathbf{Sv}=0$ can be constructed. Hence, instead of computing a single solution as in the FBA, these analyses can generate all biochemically-meaningful flux distributions based on the stoichiometric matrix. However, it is still difficult to predict the effect of genetic perturbations without resorting again to some assumptions on how cells regulate their metabolism.

To summarize, stoichiometric models with the steady-state assumption are easy to build, but their predictive power is highly dependent on the assumption of optimality and hence is very limited. Many problems are essentially caused by the lack of dynamic and regulatory information in the modeling approach [24]. Thus, this thesis focuses on kinetic models, as detailed below.

1.1.3 Kinetic Models

When detailed information on the kinetics of cellular processes is available (e.g., enzyme-catalyzed reactions, protein–DNA binding or protein–protein

interactions), kinetic models as shown in Equation 1.1 can be constructed to study dynamic properties of the system. Based on the assumed functionality of the flux vector, kinetic models can be generally divided into three categories (Figure 1.2):

(1) Mechanistically Based Models:

These models are built on biological mechanistic understanding, such as using the formulism of mass action [25] or Michaelis–Menten (MM) rate law [26], of which the former is applied to describe elementary reactions and the latter is to describe simple enzymatic reactions. However, which formula to use may become difficult to be determined *a priori*, especially for complex biochemical reactions, which involve non-elementary reactions or are catalyzed by enzymes that are not understood in sufficient detail.

(2) *Ad hoc* Models:

When detailed information on biochemical reactions is unknown or unclear, *ad hoc* black-box models, which are formulated to fit the observations, can be constructed. But these models can be highly arbitrary in formulism and structure, and involved parameter estimation may become very problematic [27]. In many cases, a canonical model could be a better option (see below).

(3) Canonical Models:

Canonical models have homogeneous structures and their individuality comes from different values of model parameters. This property keeps the model structure case-independent and simplifies the method development for model analysis and parameter estimation.

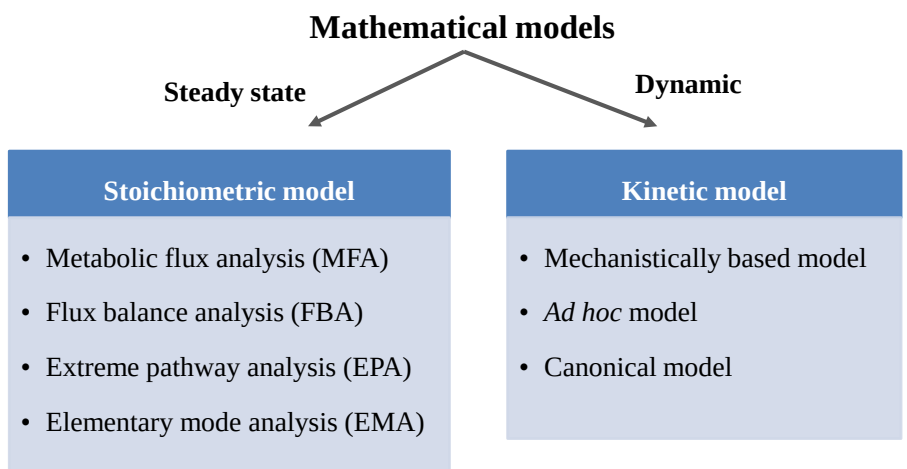


Figure 1.2. Mathematical modeling of metabolic pathways.

Among canonical models of biochemical systems, power-law models under the Biochemical Systems Theory (BST) [28,29], including Synergistic-system (S-system) and Generalized Mass Action (GMA) [30], have drawn much attention for many reasons [24]. This type of model consists of a set of differential equations, which can be generalized as:

$$\dot{X}_i = f_i(X_1, X_2, ..., X_m), \tag{1.2}$$

where X_i is the concentration of the i-th metabolite, and its change depends on some of the independent variables X_1, X_2, ..., X_m. In the S-system model, the multivariate function f_i is divided into two parts, denoting an influx (production) term and an efflux (degradation) term:

$$\dot{X}_i = v_i^+(X_1, X_2, ..., X_m) - v_i^-(X_1, X_2, ..., X_m) \tag{1.3}$$

In this case, the aggregate influx (v_i^+) and efflux (v_i^-) terms are represented by power laws:

$$\dot{X}_i = \alpha_i \prod_{j=1}^{m} X_j^{g_{ij}} - \beta_i \prod_{j=1}^{m} X_j^{h_{ij}} \tag{1.4}$$

Here, model parameters consist of rate constants α_i, β_i and kinetic orders g_{ij}, h_{ij}. The rate constants are non-negative real numbers, and the kinetic orders can take any real values, the sign of which indicates the nature of the connectivity among metabolites: positive represents a substrate or activation and negative refers to an inhibition.

Unlike the S-system, GMA formalism does not aggregate v_i into single influx and efflux terms, but here each reaction is written as a separate power-law flux, giving:

$$\dot{X}_i = \sum_{p=1}^{p_i} \left(\pm r_{ip} \prod_{j=1}^{m} X_j^{f_{ipj}} \right), \tag{1.5}$$

where again the rate constants γ_{ip} are non-negative and the kinetic orders f_{ipj} can take any real values. One can rewrite Equation 1.5 into the form of Equation 1.1 with power-law flux functions.

The formulations of the S-system and GMA models differ only at metabolic branch points (i.e., where there are multiple arrows going into or out of a node), while their other details remain the same. The S-system model reserves highly generic formalism, while the GMA model is considered to be closer to biochemical reality.

The power-law formulations are specifically designed to mimic kinetic reactions, and are sufficiently general to model metabolic pathways, as well as

other biological systems, including genetic networks [31], multi-level systems [32] and signal transduction cascades [33]. Their highly ordered mathematical structure (power-law) facilitates numerical analyses, and is able to capture any forms of non-linear behaviors (e.g., oscillation or chaos) [34,35]. As canonical models, these power-law models can be set up without much mechanistic information of the system. In addition, parameter values (i.e., rate constants and kinetic orders) directly characterize the connectivity of the metabolic pathway, as described above, and this one-to-one relationship (between kinetic parameters and structural features) facilitates parameter estimation and structure identification in a single identification step. Namely, if the knowledge of structural properties is available, it can be directly applied to determine where the corresponding parameters shall appear in the BST models. Conversely, if a parameter has been identified, its interpretation in terms of structural features is also immediate [28-30]. All the aforementioned advantages of the BST models give motivation to the major focus on this model framework in the case studies here.

1.2 Kinetic Model Construction

To construct a kinetic model, one requires the detailed information about the structure and kinetic parameters of the system, which is typically not available *a priori*. An inference method is thus desired to extract information about the structure and dynamics of the system from experimental data, and such "model building" task consists of several major phases as shown in Figure 1.3. Briefly, based on prior knowledge and time-course data, the first major phase requires structure identification to infer the topology of the metabolic network. A network graph is established using nodes to represent metabolites or other biological molecules and arrows to denote transformations between them. Following this, a suitable modeling framework, like an S-system or GMA model, is chosen to represent the system dynamics. Given the model equations, the next phase is to estimate unknown model parameters by matching model simulations with experimental observations. In the following, model invalidation can be conducted, either using information from other sources or independent experimental observations. If the model is proved to be invalid, a model refinement and new data generation will be necessary before repeating the procedure again. This process is iterative until the model is deemed to be reliable and appropriate for end-applications. For example, such model can be analyzed for the information about steady state, sensitivity and other dynamic features of the metabolic network.

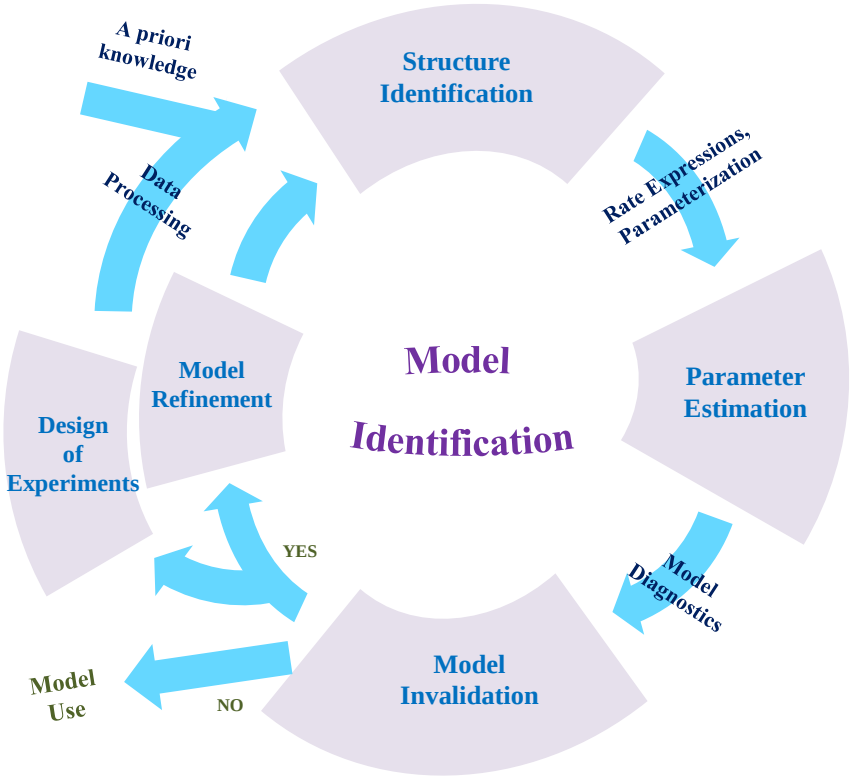


Figure 1.3. An iterative procedure of model identification.

The development of model identification methods is driven by the availability of experimental data, where different types of data require distinctly different methods. Based on many in-depth studies, the methods can be generally divided into two: forward (bottom-up) strategy and inverse (top-down) strategy. The former builds the model up by integrating "local" kinetic information on individual metabolites, enzymes and modulators, while in the latter, metabolic network topology and parameter values are directly inferred from "global" time-series data.

1.2.1 Forward (bottom-up) Strategy

Forward strategy follows a traditional reductionist approach for mathematical modeling in biology before the availability of high-throughput and/or systems-wide data. Early metabolic modeling studies were developed from “local” kinetic information. For instance, one particular enzyme, catalyzing a particular reaction within a metabolic pathway of interest, was purified and characterized one at a time to determine its optimal temperature, pH, quantified cofactors and modulators. Then this information was converted into a suitable function or rate expression such as Michaelis–Menten or Hill rate law. Once the reactions in the metabolic pathway had been identified, all the collected information would be merged into a comprehensive pathway model (e.g., see [30,36]). This model identification process benefits greatly from available databases such as KEGG [37,38], MetaCyc [39] and Brenda [40], which collect information on pathway topologies and kinetic parameters retrieved from literature. The strategy of studying these ‘local’ components (one enzymatic reaction at a time) and combining them into a more comprehensive metabolic model is known as “forward” or “bottom-up” modeling.

The advantage of this strategy lies in its straightforward nature and a direct use of available information. However, the biggest drawback is that the model built from the descriptions of individual processes seldom works as observed or expected as a whole in practice. Specially, knowledge about many constituents and processes in the model is often studied individually, where “local” data were obtained either from different organisms or from experiments conducted under

different conditions and often *in vitro* [41,42]. It is thus difficult to predict how the same constituents and processes will behave in a particular organism under the conditions of interest. Furthermore, the processes of involved model building and iterative refinements are usually labor intensive, requiring a combination of biological and computational expertise [24]. These severe drawbacks bring the next strategy to the stage.

1.2.2 Inverse (top-down) Strategy

Now, modern techniques of molecular biology are able to produce time-series data which measure the responses of a whole pathway to a stimulus, such as a change in experimental inputs or environmental conditions. In contrast to the "local" data, the appeal of such "global" data is that the measurements are taken simultaneously *in vivo* or *in vitro*, providing time-series snapshots of cellular constituents and processes. These measurements contain valuable information regarding the functional connectivities and regulations of biological networks. The information within such time-course data, however, is implicit, requiring regression analyses and estimation methods.

The inverse modeling from data is depicted in Figure 1.4. This model identification process begins with comprehensive data at a system level, which ideally consist of simultaneous time-course measurements on metabolites, gene expression or protein abundance in the same organism or cell type under identical conditions. First, there may be a need for data processing, such as a smoothing method to remove experimental noises. In the figure, power-law model

formulation (S-system or GMA model) is selected for modeling the reactions because of its advantages discussed earlier in Section 1.1.3. Thus, structure identification is integrated into the process of parameter estimation. If any prior knowledge of topology and regulation is available, it can be converted as constraints in parameter estimation, which is performed next to determine parameter values by fitting to the time-course data. Typically, the solutions are not unique but suggesting alternative network candidates that are all consistent with the provided data, so proposals for model invalidation are provided next. This iterative process of system inference is repeated until no further improvement can be made.

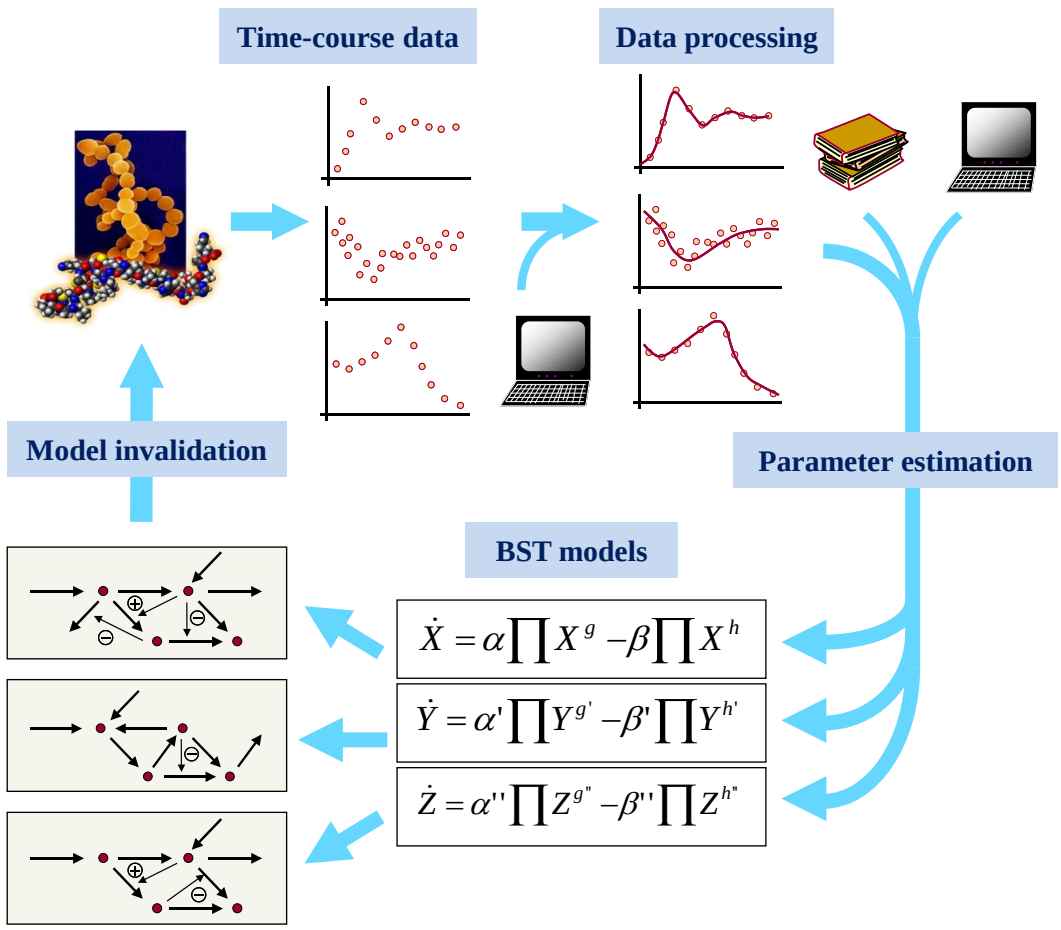


Figure 1.4. Inverse strategy of model identification.

In practice, several challenges exist in this inverse modeling process rooted from the complexity of biological systems, which will be discussed in the next chapter.

CHAPTER 2 : CHALLENGES AND OPEN PROBLEMS IN THE INVERSE MODELING

2.1 Challenges in the Inverse Modeling

The difficulties in this inverse modeling approach generally fall into one of four categories: data-related, model-related, computational and mathematical issues (see Figure 2.1). A detailed review of these challenges has been presented elsewhere [24].

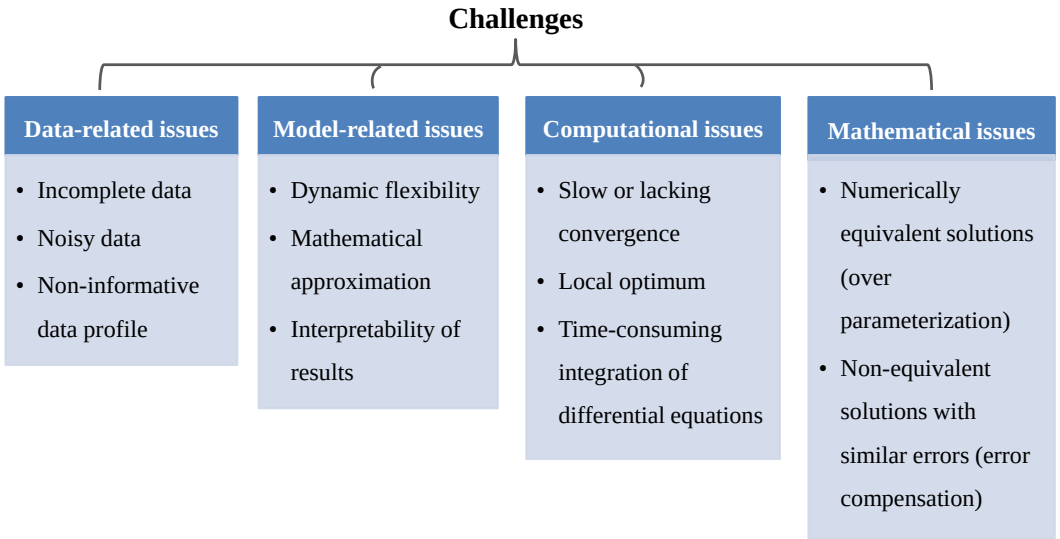


Figure 2.1. Challenges in the inverse approach of model identification.

2.1.1 Data-related Issues

There are different types of data available for model building. To characterize steady-state systems, isotopic labeling experiments have been combined with

metabolic flux analysis, allowing for a reliable estimation of fluxes, especially for some unmeasured intracellular fluxes [43-45]. Time-series measurements of metabolite concentrations can be made *in vivo* or *in vitro* by current techniques, such as Nuclear Magnetic Resonance (NMR) [46,47], Mass Spectrometry (MS) [48,49] and High Performance Liquid Chromatography (HPLC) [50,51]. NMR is more commonly used for online *in vivo* measurement, coupled with isotopic labeling, e.g. C^{13} for glycolytic metabolites and P^{31} for ATP, Pi. The involved experimental procedure includes sample preparation and on-line NMR measurement [52].

However, the datasets from these experimental measurements are seldom complete due to two roadblocks particular in biology: complexity and technology. First, a metabolic network typically involves a large number of metabolites with complex connectivity, which means that the complete measurement of all relevant metabolites is practically not feasible. These problems are especially severe for the intermediate species, which may be very difficult to measure explicitly. Second, in order to capture the dynamic behaviors of the metabolites, time-course data must be measured accurately and frequently enough, which often challenges the limit of current available techniques. In practice, data collection could be missing at certain time points because of various reasons (e.g., human error). The issue of this missing time-points can be partly addressed by standard interpolation, and in a few instances, it may be possible to obtain the missing metabolite measurements by analyzing the left null space of stoichiometric matrix to generate sets of metabolites whose total weighted concentrations are time

invariant [53]. However, a complete loss of data for certain metabolites poses a much more challenging problem in parameter estimation, which requires more sophisticated methods to bridge the left gap. This problem will be tackled in Chapters 3 and 4.

Even when data are complete, they are usually noisy due to technical or human reasons. To this end, data smoothing methods, such as splines [54-56], polynomial fitting [57], filters [58] and artificial neural networks (ANNs) [59,60], can be employed to alleviate the problems associated with measurement noise. Although the methods of splines are easy to be implemented, they may produce artificial fluctuations in the smoothened curves when the data are very noisy. On the other hand, polynomial fitting is an efficient and widely applied method, but additional care needs to be taken to avoid over-fitting problems. Common filters such as Kalman, Savitzky-Golay and Whittaker filters have also been used [58]. For example, Vilela and co-workers [61] had presented a Whittaker-Eilers smoother and its implemented software AutoSmoother, in which the optimization criterion is defined as Renyi's second-order entropy of the cross-validation error. Almeida et al.[59] applied ANNs to biochemical time-series data, showing the great promise of this method. The interpolating functions obtained from ANNs are universal and flexible, but may lead to artifacts in the slope approximation, e.g., resulting in an undesirable offset in the smoothed data.

Aside from frequency and accuracy of measurements, another data-related problem is due to "non-informative" experiments, e.g., some metabolite time-profiles are co-linear or constant. Such co-linearity may cause ill conditioning of

the estimation process, a problem known as parameter identifiability issue [62]. There exist methods through which the lack of complete parameter identifiability can be assessed, even prior to parameter estimation [63,64].

2.1.2 Model-related Issues

The inverse problem asks for an “ideal” mathematical model to be capable of capturing all possible nonlinear dynamics of the system while keeping the involved mathematics relatively simple. As introduced in Section 1.1.3, the feasible model candidates include a large variety of structures and mathematical formulations. Some models are mechanistically formulated, some are only meant for data fitting regardless of model structure and others try to achieve a balance between the aforementioned two.

Mechanistic models are commonly used in modeling chemical reactions, and have also been applied to describe biological phenomena. In practice, this approach may not always be the best choice due to two reasons. On the one hand, the exact mechanisms of the targeted biochemical reactions are seldom known completely, so that the potential model candidates may include a number of models with different mechanistic formulations. On the other, time-course experimental data are often not sufficient and accurate enough to discern among those candidates. As a result, it is more prudent to adopt a generic approach, meeting the demands including dynamic flexibility to capture important features of time-course data, simplicity of mathematical approximation to represent the system, and interpretability of the parameter estimation results for biological

meanings behind. To this end, the power-law representations under the BST, as described above, are especially useful to overcome some of the model-related issues. Chou et al. [24] listed the common metabolic models used for testing method algorithms, including a three-variable cascaded pathway [65,66], a four variables didactic system [67], a four-variable branched pathway [60,66], a five variables gene regulatory network [68], a five-variable ethanol fermentation model [69], the five-variable metabolism model in *E. coli* [70], the anaerobic fermentation pathway in *S. cerevisiae* (five dependent variables and eight independent variables) [71-74], the five-variable glycolysis pathway in *S. cerevisiae* [66], the six-variable glycolysis pathway in *L. lactis* [75-78] and the eight-variable trehalose pathway in *S. cerevisiae* [79,80].

2.1.3 Computational Issues

One of computational challenges in the inverse modeling lies in the expensive numerical computation for model solutions. For ODE models shown above, numerical integration can be extremely computationally expensive to perform during estimation. One study showed that such numerical integrations consumed the majority of computational resources during the parameter estimation, up to 95% [60]. In another study, the application of standard parameter estimation methods (e.g., least square or maximum likelihood) to an S-system model encountered numerical integration problems due to ODE stiffness (a numerical difficulty caused by large differences in time scales among simulations), leading to non-convergence of the estimation results [66]. While such stiffness can genuinely

arise due to a large time scale separation of the reaction kinetics in the real system, stiff ODEs could also result from unrealistic combinations of parameter values during the parameter optimization procedure, especially when a global optimizer is used. The parameter estimation of ODE models using power-law kinetics is particularly prone to stiffness problem since many of the unknown parameters are the exponents of the concentrations. To circumvent this computationally-costly integration of ODE models, several methods have been proposed, such as decoupling [30,60], ODE decomposition [31,81] and collocation methods [65]. Some of these methods form the basis for the present thesis.

Furthermore, as the typical parameter estimation is formulated as a minimization of model prediction error, complicated error function surfaces can result in a slow convergence toward global minimum or convergence to local minima. In addition, the parameterization of kinetic ODE models often lead to a combinatorial increase of unknown parameters along with the increasing number of metabolites, resulting in a large-scale optimization problem. Overcoming these difficulties calls for powerful global optimization tools [31,60,82] and sufficient constraints for parameter search space [30,83].

To reduce the computational requirements of performing parameter estimation, incremental estimation methods have been proposed [77,84]. In these methods, dynamic metabolic fluxes are first estimated and the parameter estimation is subsequently done one flux at a time. Such incremental identification approach generally has the advantages of low sub-problem complexity, low computational effort, flexible use of physically motivated equations for each flux, and ease of

validation of flux equations [85]. Nevertheless, more work is still required to make the approach more efficient for metabolic network modeling.

2.1.4 Mathematical Issues

An often-ignored problem in parameter estimation is mathematical redundancy in some models. Even after more than 100 publications in the applications of BST modeling to biochemical networks, the parameter estimation remains a bottlenecking step. Different estimation techniques often produce widely different parameter estimates and these parameters could fit experimental data equally well [86]. One possible cause lies in model formulation, where there could be a case of over-parameterization. For instance, if two parameters p and q always enter an equation in the same combination as $(p+q)$, then their individual values cannot be identified. In essence, the difficulty in identifying p and q individually results from the fact that perturbations in each parameter will cause the same changes in the system outputs, and thus they cannot be differentiated from looking at the output measurements.

It may also happen that non-equivalent solutions exhibit similar residual errors. In the context of power-law formulas, error compensations can occur within or between metabolic fluxes, producing different rate constants and kinetic orders with similar model prediction errors. Such error compensations may be caused by degrees of freedom in the inverse problem. For example, when the number of metabolites is smaller than the number of reactions, there exist many flux values that satisfy Equation 1.1, a common circumstance in metabolic networks. Since

some metabolites in the pathway can participate in more than one reaction, e.g. the pathway usually has branched or reversible reactions, the issues associated with underdetermined systems will be very likely encountered. For example, the GMA model of the three-variable cascaded pathway, introduced in Section 2.1.2, has 2 degrees of freedom, and the 5-variable glycolysis pathway in *S. cerevisiae* has 3 degrees of freedom. This kind of issue will be tackled in Chapters 4 and 5.

These are other contributors of parameter identifiability, aside from the aforementioned data issues. The situation can be much improved by performing more and better experiments that cover wide ranges of input variations. *A priori* kinetic information on individual reactions can also help in this case and should always be incorporated if available [87].

In response to the four issues discussed above, many studies have been working on the solutions. A representative collection of these studies will be reviewed in the next section.

2.2 Support Algorithms for the Inverse Approach

Many advanced techniques for the inverse approach of model identification have been developed and the representative support algorithms are listed historically in Table 2.1.

Table 2.1. A historical listing of the representative support algorithms for the inverse modeling approach.

Authors	Year	Methods	Features	Model
Bonvin and Rippin **[88]**	1990	Target factor analysis	Stoichiometry check	—
Chevalier et al.**[89]**	1993	Evaluation of stationary-state Jacobian matrix elements	Structure identification	BST
Arkin and Ross **[90]**	1995	Correlation metric construction: analysis of a time-lagged multivariate correlation function	Structure identification	Mass action
Tominaga et al.**[91]**	2000	Genetic algorithm	Parameter estimation	S-system
Samoilov et al.**[92]**	2001	Entropy metric construction, entropy reduction method	Structure identification	Mass action
Maki et al.**[93]**	2002	Step-by-step strategy (decomposition method)	Genetic network inference	S-system
Vance et al.**[94]**	2002	Direct observation for causal connectivities	Structure identification	MM, BST
Kikuchi et al.**[95]**	2003	Penalty on small kinetic orders, genetic algorithm with simplex crossover method	Kinetic network inference	S-system
Veflingstad et al.**[96]**	2004	Multivariate linear regression on data	Data processing, Parameter constraining	BST
Crampin et al.**[97]**	2004	“general-to-specific” approach	Kinetic network inference	Mass action
Voit and	2004	Decoupling method, ANN smoothing	Parameter	S-system

Almeida **[60]**			estimation	
Katare et al.**[98]**	2004	Particle swarm optimizer with Levenberg-Marquardt method	Parameter estimation	—
Kimura et al.**[99]**	2005	ODE decomposition strategy, cooperative coevolutionary algorithm.	Kinetic network inference	S-system
Tsai and Wang **[65]**	2005	Data collocation, hybrid differential evolution	Parameter estimation	S-system
Tucker and Moulton **[100]**	2005	Interval analysis for parameter reconstruction	Parameter constraining	S-system
Marino and Voit **[101]**	2006	"Simple-to-general" approach, gradient-based optimization	Model generation, model fitting, model selection	S-system
Marquardt et al.**[102]**	2006	Incremental identification	Kinetic network inference	Mass action
Polisetty et al.**[73]**	2006	Branch-and-reduce method	Parameter estimation	GMA
Daisuke and Horton **[103]**	2006	Distributed genetic algorithm, scale-free network	Kinetic network inference	S-system
Cho et al.**[104]**	2006	S-trees representation, genetic programming	Biochemical network inference	S-system
Kutalik et al. **[105]**	2007	Newton-flow optimization	Parameter estimation	S-system
Tucker et al.**[106]**	2007	Interval analysis with constraint propagation	Parameter constraining	GMA
Noman and Iba **[107]**	2007	Information criteria-based fitness evaluation, differential evolution	Genetic network inference	S-system
Gonzalez et al.**[108]**	2007	Simulated annealing algorithm	Kinetic network inference	S-system
Vilela et al.**[109]**	2008	Eigenvector optimization, parameter pruning	Kinetic network inference	S-system
Goel et al.**[77]**	2008	Dynamic flux estimation	Kinetic network inference	GMA
Zuniga et al.**[110]**	2008	Ant colony optimization algorithm	Parameter estimation	S-system

Machina et al.**[111]**	2010	Automated piecewise power-law modeling	Data processing, parameter estimation	Piecewise power-law
Zhan and Yeung **[112]**	2011	Spline approximation, linear and nonlinear programming	Parameter estimation	Mass action, MM

2.2.1 Methods of Data Processing and Model-free Structure Identification

The information about metabolic network topology and constraints for parameter search space can be inferred using data processing methods, where various methods have been developed for data with different characteristics. For instance, one method relied on the transient measurements of a metabolic system after small perturbations from steady state. In this case, the system behavior can be approximated in a linear fashion. Network connectivity was then obtained by determining the Jacobian matrix from experimental data [96,113].

Vance and co-workers [114] proposed an alternative strategy for structure deduction from direct observations of time profiles by perturbing different components in the network. This approach involved an interpretation of the profile shapes, and the observable features regarding the responses of unperturbed components can unveil the network connectivity. For example, the extreme values of the unperturbed components in response to the perturbation reveal the topological distances among them, and the initial slopes of the time courses reflect whether the components are directly affected by the perturbed component. Compared with the methods based on the Jacobian matrix above, this method has

lower limitation of experiments, in which the perturbations can be done in arbitrary amplitude at different locations in the network.

However, this strategy may become extremely difficult for the applications to larger and more complicated systems, where complex relationships between metabolites can hardly be extracted from the simple interpretation of time profiles. In this case, correlation-based approaches (e.g., correlation metric construction and entropy metric construction) [92,115,116] may be more helpful, especially when multiple datasets covering large variations are available. Once preliminary stoichiometry of the network has been identified, Target Factor Analysis (TFA) can be applied to test each candidate for compatibility with experimental data [117,118].

2.2.2 Methods of Model-based Structure Identification

Without a completely known network topology, identification of network structure and parameter estimation can be performed simultaneously using the aforementioned power-law model formulism. One approach, coined “simple-to-general”, starts with the simplest reasonable model and gradually increases its complexity until no further improvement in the minimum modeling error function is made. For example, Marino and Voit [101] began with a simple S-system model and gradually increased the model complexity, which enabled model generation, model fitting and model selection. On the other hand, “general-to-specific” modeling initiates a full symbolic model with all parameters unknown and eliminates reactions until the model prediction error becomes unacceptable

[119]. Crampin et al. [97,120] developed global nonlinear models using elementary reactions as a basis set, and applied the "general-to-specific" and "simple-to-general" approaches to infer reaction mechanisms from time-series data. In their case studies, the results indicated that the former approach generally outperformed the later.

Generally speaking, these model-based methods are feasible only for relatively small systems, where building an all-encompassing model and finding the optimization solution is tractable. However, the estimation task can grow very quickly with the system size (i.e., the number of variables and parameters) and often suffers from a "combinatorial explosion", posing significant challenges in finding the optimal solution. Fortunately, biology naturally offers a beneficial feature that the connections within genomic, proteomic and metabolic networks are generally sparse. Studies show that the majority of metabolites are directly involved in fewer than four or five processes. For example, Thieffry et al. [121] characterized the transcriptional regulation in *E. coli* and found out that its mean connectivity fell between 2 and 3, which presents a rather loosely interconnected structure. More comprehensively, Jeong et al. [122] have examined the core metabolic network topologies of 43 different organisms representing all three domains of life (Archae, Bacterium and Eukaryote). Built from the existing databases, their study again shows that scale-free properties can describe all the chosen organisms and the average number of incoming/outgoing links per node for each organism is generally lower than 4 or 5. Furthermore, the large-scale organization of interactions among these cellular constituents appears to be

composed of small subnetworks that are loosely interconnected. Therefore, by limiting the number of structural connections within the network, the identification task can be significantly simplified.

Benefited from this general observation, parameter-pruning methods could be used to remove unlikely connections. For example, one can define a threshold for the minimal value of kinetic parameters, below which the parameters are set to be zero and thus are pruned [60,109]. Researchers have extended this pruning strategy by adding a penalty term associated with kinetic parameters in the cost function, aiming to penalize the small-valued parameters that have a negligible effect on the system dynamics [95,99,123]. However, a common challenge arising along with the introduction of the additional penalty term is that the corresponding weighted coefficient in the penalty term must be tuned carefully, since the weightings can greatly affect the estimation results.

In summary, the existing strategies of structure identification can be categorized into two groups, namely, model-free and model-based approaches (as shown in Figure 2.2), each of which has their own strengths and weaknesses. The latter can take advantage of the assumed model, such that data requirement (quantity and quality) is not as high as the former. However, proposing a model may also introduce “bias”, in which the model enforces some constraints on the set of feasible behavior that the *in silico* system could produce. For structure identification, some model-free data processing can be performed prior to model-based estimation to simplify the problem beforehand (i.e. to alleviate

combinatorial explosion). Therefore, a combination of the two approaches shall be suggested to give a more powerful identification strategy.

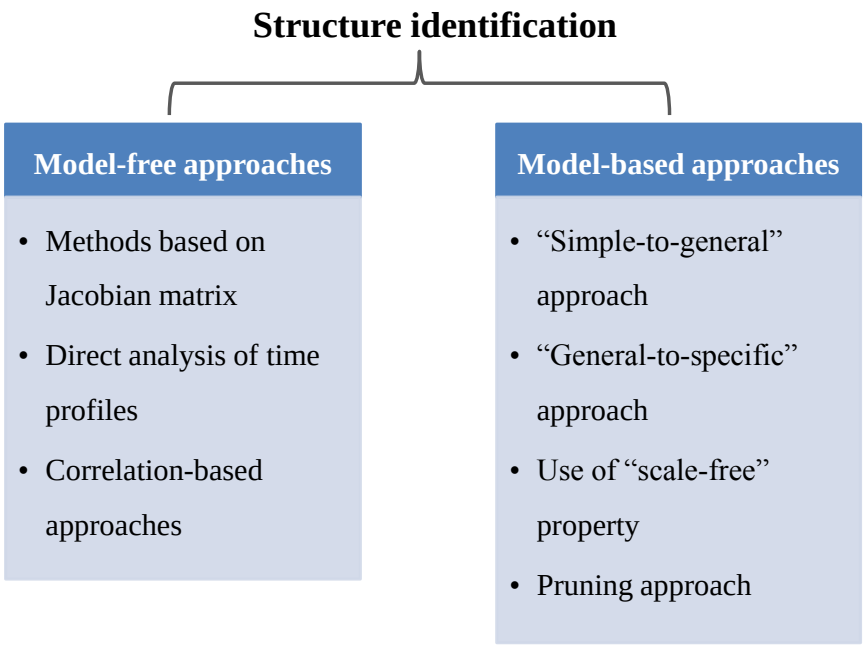


Figure 2.2. Structure identification strategies.

2.2.3 Methods of Circumventing the Integration of Coupled Differential Equations

As can be seen in Section 2.1.3, the numerical integration of coupled differential equations often requires very high computational efforts, driving the estimation task unachievable for some large-scale models. Alternative formulations have been proposed to avoid these integrations either partially or completely. In 2002, Maki et al.[93] proposed a strategy to integrate each differential equation of the ODE model one at a time, and therefore decomposed

the coupled ODE system into independent single ODEs. During the integration of an ODE, other states (metabolites) were treated as external inputs, whose values were interpolated from experimental data. The computational effort was reduced, but could still be expensive especially for large-scale model identification.

Another decoupling method was proposed by fitting the ODE model to the slopes of time-concentration data directly, thereby avoiding the ODE integrations completely and furthermore decoupling the ODEs into algebraic equations [60]. In a similar fashion, instead of fitting slopes, Tsai et al.[65] proposed a collocation method to approximate the dynamic profiles of the measured data at sampling points. In addition, Zhan and Yeung [112] combined the spline theory with (non)linear programming to remove the need for ODE solvers in the identification process. However, a big drawback is that the estimation results may become inaccurate if the measurement data are very noisy, since the mass balance is only satisfied at discrete time points (as no integration is performed between time points).

2.2.4 Methods of Constraining the Parameter Search Space

The viable range of kinetic parameters can be bounded by various types of constraints from mass balance, thermodynamics (e.g., effective reversibility or irreversibility of reactions), enzyme or transporter capacities (e.g., maximal uptake or reaction rates v_{max}) and so forth [124]. For example, several studies have provided some information about the maximal values of metabolic fluxes in some specific organisms, such as v_{max} = 4.698×10^5 mM/min in central carbon

metabolism of *Escherichia coli* [125], v_{max} = 384.2 mM/min in citric acid production of mold *Aspergillus niger* [74], v_{max} = 3440 mM/min in anaerobic fermentation of *Saccharomyces cerevisiae* [72] and v_{max} = 231.0 mM/min in purine metabolism of human cell [126]. Provided the increasing amount of information on network structure and dynamics, feasible phenotypes and their associated parameters could be further specified. For instance, under the power-law representations, prior knowledge about the network structure and regulation can be immediately interpreted as bounds of certain parameter values, because of the unique mapping from the structural features onto the model parameters. Generally, rate constants are non-negative (i.e., reactions are written as irreversible) and real-valued kinetic orders lie typically between -1 and +2 [66].

Some other studies on reducing the parameter search space are summarized below. Kutalik et al.[105] presented a parameter estimation method using a Newton-flow analysis and constructed a one-dimensional basin of attraction with true optimum contained, which significantly reduced the parameter search space. Tucker and Moulton [100] developed a parameter reconstruction method based on interval analysis, enabling an exhaustive search of the entire parameter space within a finite number of steps. This method attempted to solve the problem in a deterministic way through a pruning scheme based on a Boolean function, instead of recasting the parameter reconstruction as a global minimization problem. Tucker et al. [106] also used interval analysis in combination with constraint propagation to obtain the viable range of parameter values, which, in particular, is well suited to parameter estimation for the GMA models.

2.2.5 Methods of Incremental Model Identification

The majority of existing model identification methods, including those mentioned thus far, involve simultaneous (single-step) parameter estimation, where model prediction errors are minimized over the entire parameter space. This approach relies on efficient and robust global optimization methods (see the next section). However, as discussed in Section 2.1, the problem of obtaining the best-fit parameter estimates is ill-posed due to the issues related with data informativeness, problem formulation and parameter correlation, all of which contribute to the lack of complete parameter identifiability. Not to mention, finding the global minimum of model residuals over highly multidimensional parameter space is challenging and can become prohibitively expensive to perform on a computer workstation, even for tens of parameters. These factors motivate the development and use of an incremental identification approach.

In incremental identification, the estimation problem is decomposed into a sequence of sub-problems. For the parameter estimation of kinetic ODE model given in Equation 1.1, the model fitting to data is done incrementally: (1) obtain the rate of change in species concentrations, (2) estimate the reaction rates or fluxes, and (3) perform parameter regression for each flux. Such estimation has been applied to the modeling of complex homogeneous reaction systems [102] and to the GMA models (known as Dynamic Flux Estimation (DFE)) [77]. Recently, Machina et al. [111] extended the DFE method by adopting piecewise power-law functions, which offered an effective solution to produce an almost unbiased representation of time-series data.

The existing incremental estimation methods generally have the advantages of low computational effort, low sub-problem complexity, flexible use of physically motivated equations for each flux, and ease of validation for flux equations [85]. However, available methods typically assume that the number of species is at least equal to the number of reaction fluxes, such that the estimation of fluxes from the rate of change of species concentrations can produce a unique solution. However, in the typical metabolic networks, the number of measured metabolites is smaller than that of fluxes, as a metabolite usually participates in more than one reaction. Thus, a generalization of this incremental estimation approach is urgently needed (see Chapters 4 and 5 for further details).

2.3 Optimization Algorithms

The aforementioned identification strategies typically rely on finding a global optimal solution to a nonlinear programming problem. Hence, the efficacy of any strategies relies heavily on the choice of optimization algorithms. These algorithms could be generally categorized into three groups: deterministic, stochastic search and hybrid optimizations, as summarized in Figure 2.3.

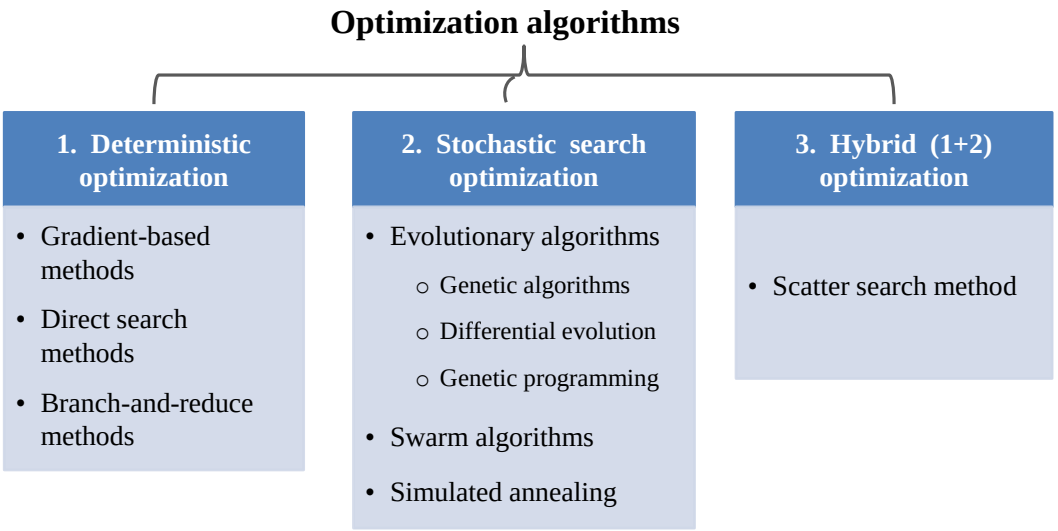


Figure 2.3. Optimization algorithms: deterministic, stochastic and hybrid optimizations.

2.3.1 Deterministic Optimization Algorithm

If the gradients of objective functions can be (cheaply) computed, the most common optimization algorithm for parameter estimation is a gradient-based approach, where the search for optimality corresponds to finding the parameter values for which the gradient vanishes [127]. Many methods of this type have

been applied to metabolic network modeling. For example, Marino and Voit [101] developed an automated information extraction procedure involving gradient-based optimization methods, and this procedure combined model generation, parameter estimation (model fitting) and model selection using S-system models. As mentioned earlier, Kutalik et al.[105] employed a Newton-flow optimization for S-system parameter estimation and constructed a one-dimensional basin of attraction where the true optimum resides.

Of course, gradient-based methods cannot be applied to the cases where the objective functions or associated derivatives are discontinuous. More importantly, the parameter estimation of nonlinear ODE models typically encounters non-convex objective function surface with many local minima. The efficacy of gradient-based search methods often depends on the starting points of the optimization in order to converge to the global optimum [128], i.e., one should start with the initial parameter guesses close enough to the global solution. Branch-and-bound strategy could be useful in avoiding the premature convergence to local optimum solutions [129]. For power-law models, Polisetty et al.[73] introduced a branch-and-reduce method and formulated a convex optimization problem. However, the major drawback of this method is the high computational requirement.

2.3.2 Stochastic Search Optimization Algorithm

This group of optimization methods introduce stochasticity (randomness) into the optimization algorithm in order to prevent the search process from getting

trapped into local minima [130]. Examples of stochastic search optimization methods include Evolutionary Algorithms, Swarm Algorithms and Simulated Annealing, which have been widely applied in the model identification of biological systems.

(1) Evolutionary Algorithms

Evolutionary Algorithms (EAs) were developed for generic population-based meta-heuristic optimization. Inspired by the process of natural evolution, EA incorporates biologically motivated mechanisms in the optimization process, including reproduction, mutation, recombination and selection. Some examples of widely used algorithms in this group include Genetic Algorithms, Differential Evolution and Genetic Programming. These methods only differ in the details of the implementations of the evolutionary processes.

Genetic Algorithms (GAs) [131] have been routinely used in the parameter estimation of power-law models. For instance, Tominaga et al.[91] implemented a simple version of GA and an S-system formulation to predict structure and dynamics of a simple oscillation system and a gene expression network. However, the proposed method could only predict a small number of parameters and its convergence was slow. Responding to these drawbacks, Kikuchi et al.[95] extended the method by adding a structure-related penalty term and by employing a novel crossover method and a gradual optimization strategy. However, the computational cost was still somewhat high due to the costly numerical integrations of coupled ODE systems. In addition, several studies have been

conducted to improve the efficiency of GA when using time-course data in the inverse modeling of power-law models. For example, Okamoto et al.[132] incorporated GA into a modified Powell method and case studies showed that this procedure can locate the global minimum with considerably fast convergence. Ueda et al. [133,134] proposed an efficient optimization technique for S-system models based on real-coded GA using unimodal normal distribution crossover and minimal generation gap. In addition, Daisuke and Horton [103] developed distributed GA to estimate parameters of S-system models with scale-free properties, and Ho et al.[135] proposed another modified GA with intelligent crossover based on an orthogonal experimental design to efficiently infer genetic networks in a two-stage manner.

Differential Evolution (DE) is another type of commonly used evolutionary algorithms [136,137]. Using DE coupled with a hill-climbing local search, Noman and Iba [138-140] introduced a novel fitness evaluation based on information criteria to infer gene regulatory networks, instead of the conventional fitness defined by least-squared errors. Tsai and Wang [65] applied hybrid DE to obtain a starting point for gradient-based optimization methods and used a collocation method to convert ODEs into algebraic equations. Furthermore, they implemented a multi-objective optimization method with the hybrid DE to estimate the parameters of S-system models [141].

Genetic Programming (GP) [142] adopts a tree-based internal data structure to represent computer programs or mathematical expressions, providing a general technique for identifying metabolic pathway models from time-course data. Koza

et al. [143] applied the GP optimization to construct the topology of metabolic pathways and identify the rates of involved chemical reactions. Sugimoto et al.[144] introduced numerical mutations into the conventional GP procedure and added a penalty term to the cost function, enabling a simultaneous search for the network topology and its parameters without the complete knowledge of biochemical reaction mechanisms. On the other hand, without adding any penalty terms, Cho et al.[104] showed that one can still identify the network structure and the involved parameters at the same time by proposing a new S-tree representation for the network models, which functioned efficiently with the S-tree based GP method.

Generally speaking, EAs perform quite well in the most parameter estimation problems because of their generality, without any requirements on fitness landscape and on differentiability of objective functions. In addition, the framework is amenable to parallel implementation for use in a high performance computing cluster. However, by mimicking the evolution process, these algorithms often experience a slow convergence to the global optimal solution. Some comparisons among the aforementioned techniques have been presented by applying them to benchmark nonlinear programming problems (subject to nonlinear differential algebraic constraints) [145,146]. The studies found that gradient-based local optimization methods often failed to reach satisfactory solutions. Although some evolutionary algorithms like the real-valued GA and DE were capable to handle complex and multi-modal search space, additional

fine-tuning of the corresponding algorithm settings was still crucial for the success of the optimization.

(2) Swarm Algorithms

Swarm Algorithms were developed to replicate the collective behaviors of decentralized and self-organized natural systems, such as those found in an ant colony or a school of fishes. Some methods belonging to this classification include Ant Colony Optimization (ACO) and Particle Swarm Optimization (PSO), of which the former is motivated by ant foraging through pheromone communication to form paths [147] and the latter is modeled on how a swarm of particles move in the search space based on the information shared among swarm members [148]. Zuniga et al.[110] successfully applied the ACO strategy for network inference using S-system models, while Naval et al.[149] used modified PSO to infer the kinetic parameters of S-system and GMA models. Like the aforementioned EAs, the computational requirement for these algorithms is high and convergence is often slow. Similarly, the computation of objective functions can be straightforwardly parallelized, if desired.

(3) Simulated Annealing

Simulated Annealing (SA) [150] obtained its name and inspiration from annealing process in Metallurgy, a process of heating and controlled cooling to achieve minimal internal energy state of materials. By analogy with this physical process, the SA algorithm attempts to reach a “state with the minimal energy” (minimum of objective function) from an arbitrary “initial state” (a point in

parameter search space) by adjusting “cooling-down schedule” (a strategy of sample reproduction and selection). Ceric and Kurtanjek [151] applied SA to partial re-estimation for the kinetic parameters of the central metabolism in *Escherichia coli*, and the built model could correctly predict oscillatory responses upon glucose impulses. Gonzalez et al.[108] also applied SA to estimate the kinetic parameters of S-system models from time-course data, and the efficiency of this method has been proved in the applications of artificial and actual metabolic networks.

The convergence of the SA algorithm can be rigorously proven and SA can perform like a global stochastic optimization technique at high “temperatures” and gradually behave like a local optimization technique at low “temperatures”. Unlike the aforementioned stochastic algorithms, this method generally involves time-costly computations that are not easy to be parallelized.

2.3.3 Hybrid Optimization Algorithm

Deterministic methods are generally faster than stochastic search methods, but run a high risk of getting trapped in local minima especially for non-convex problems. On the other hand, stochastic methods can significantly increase the chance of finding the global optimal solution by vastly exploring the parameter space, but often at the cost of great computational effort and slow convergence. Consequently, new methods have been created by combining both strategies in order to arrive at the global optimal solution more efficiently. In these methods, stochastic searches are initially applied to isolate local regions in the parameter

space in which the global optimal solution may lie and subsequently, gradient-based methods are performed using the initial parameter guesses from the stochastic search optimization.

Katare et al.[98] proposed a hybrid algorithm for parameter estimation, combining population-based stochastic PSO with a Levenberg-Marquardt optimizer. Their results showed that this hybrid method was more effective than the PSO alone in finding the global optima for six benchmark problems. In their another publication, Katare et al. [152] used GA to identify the promising regions of parameter search space and further explored these regions locally by a modified Levenberg–Marquardt method, again, achieving the global optimum much faster than the GA alone. In addition, a number of studies have demonstrated the advantage of hybrid strategy over each individual approach, e.g., combining evolutionary strategy [153] with other deterministic search methods, like Gauss-Newton [154], Nelder-Mead [155] and Levenberg-Marquardt [156] methods.

A notable hybrid optimizer is a Scatter Search method (SSm), a population-based meta-heuristic global optimization method incorporating stochastic and deterministic strategies [157-159]. Importantly, this method offers an automated switch between global search phases and local intensification phases with a number of local solvers available in the toolbox, so that it is very effective in solving multi-minima optimization problems. These advantages motivate the use of the SSm as the global optimizer in this thesis.

2.4 Open Issues and Thesis Scope

Among the aforementioned challenges in the inverse modeling of metabolic networks, previous studies mostly addressed the first three classes: data-related, model-related and computational issues, but mathematical issues have not been adequately investigated. However, more efforts are still required for tackling the unresolved difficulties in all of these four categories, which are summarized below and in Table 2.2.

The use of canonical models, like power laws, often leads to a large-scale parameter estimation problem and stiff ODEs, and consequently to a computationally intractable estimation. As described above, methods exist that avoid the integration of ODEs, such as the decoupling method, which can alleviate the cost associated with the numerical integration of stiff ODEs. However, these methods require the complete measurements of all species in the pathways, and thus the practical applications could be very limited. In this thesis, I have developed a new method to circumvent the issue of missing metabolic time profiles in the application of decoupling strategy, as to be described in Chapter 3.

Thus far, the parameter identifiability issue has not been directly tackled during the parameter estimation. The lack of complete parameter identifiability simply means that there exists no unique solution to the estimation problem. As discussed previously, such identifiability issue arises from the existence of degrees of freedom in the estimation problem due to, for example, noise in the data, parameter correlation or model formulation. The parameter correlation can

also affect the optimization process negatively, where the dimension of the parameter search space is unnecessarily large. Chapters 4 and 5 describe two attempts to address the identifiability issue directly using an incremental estimation approach.

Two incremental estimation methods have been developed to handle the degrees of freedom that arise from having more reactions or fluxes than species, a common circumstance in metabolic pathways. Existing incremental approach cannot be applied in this situation. Such degrees of freedom mean that there could be (infinitely) many flux combinations ($\mathbf{v}(t)$) that match data on the measured rate of change of species concentrations ($\dot{\mathbf{X}}_m(t)$). There are two implications from this fact. First, there are two groups of parameters, defined according to how they enter the flux functions, where one group can be set independently and the rest can be computed from the first group using the relationship in Equation 1.1. In other words, the parameter estimation search space can be reduced to only over the first group of parameters, and this becomes the premise of a new highly-efficient parameter estimation method. Second, if one treats any feasible flux combinations matching $\dot{\mathbf{X}}_m(t)$ as equivalent, then one can create an ensemble of kinetic models that agree with the given data. Thus, in the second incremental estimation approach, the method is designed to provide the ensemble of all biologically meaningful kinetic models of the metabolic networks, given time-series measurements of metabolite concentrations and realistic bounds on the parameter values.

Note that model identification consists of several major phases, including data processing, structure identification, parameter estimation, model invalidation and experimental design (as shown in Figure 1.3). All these steps are challenging, especially for biological systems. Among these, the parameter estimation is the central topic of the present thesis. The development of parameter estimation algorithms could also benefit model-based structure identification, especially under the BST model framework (as introduced in Section 2.2.2). While this thesis provides solutions to some important and challenging problems that arise in the kinetic modeling of metabolic pathways, other challenges may still remain. Chapter 6 provides a short discussion of the remaining challenges and some ideas on how to tackle them based on the findings of the work presented in this thesis.

Table 2.2. Challenges, solutions and my work in the inverse modeling approach.

Challenges		Solutions	Method 1 (Chapter 3)	Method 2 (Chapter 4)	Method 3 (Chapter 5)
Data-related	Incomplete data	Interpolation (missing points)			
		Sophisticated methods (missing metabolites)	★		
	Noisy data	Smoothing methods	✓	✓	✓
	Non-informative	Identifiability analysis; Inference methods	✓		
Model-related	Unknown mechanism of reactions	Canonical models (BST)	✓	✓	
	Different dynamic formulations	Hybrid models			✓
	Unknown structure	Structure identification			
Computational	Slow or lacking convergence; Local optima	Global optimization	✓	✓	✓
		Constrained parameter space		★	✓
	Stiff ODEs	Decoupling method	✓	✓	✓
		Decomposition method	★		
		Collocation method			
Mathematical	Over parameterization; Error compensation	Incremental identification		★	★
		Ensemble modeling			★
		Additional kinetic information			

Tick: involved methods; Star: key focuses.

CHAPTER 3 : TWO-PHASE DYNAMIC DECOUPLING METHOD

3.1 Summary

Time-series measurements of metabolite concentrations have become increasingly more common, providing data for building kinetic models of metabolic networks using ODEs. In practice, however, such time-course data are usually incomplete and noisy, and the estimation of kinetic parameters from these data is challenging. Practical limitations due to data and computational aspects, such as solving stiff ODEs and finding global optimal solution to the estimation problem, give motivations to develop a new estimation procedure that can circumvent some of these constraints.

In this chapter, an iterative parameter estimation method is proposed that combines and iterates between two estimation phases. One phase involves the decoupling method, in which a subset of model parameters associated with measured metabolites is estimated using the minimization of slope errors. The other phase follows, in which the ODE model is solved one equation at a time and the remaining model parameters are obtained by minimizing concentration errors. The performance of this two-phase method was tested on a generic branched metabolic pathway, an *E. coli* metabolism model and the glycolytic pathway of *Lactococcus lactis*. The results show that the method is efficient in obtaining accurate parameter estimates, even when some information is missing.

3.2 Method

3.2.1 Decoupling Method

In order to circumvent expensive computational efforts in solving coupled ODEs, a method was proposed previously by fitting the right hand side of the ODE model in $\dot{\mathbf{X}}(t) = f\left(\mathbf{X}(t);\mathbf{p}\right)$ to the slopes of concentration data directly, thereby decoupling the ODEs [28,29,60]:

$$\begin{aligned}
S_1(t_1) &\approx \dot{X}_1(t_1) = f_1\left(X_1(t_1), X_2(t_1), \ldots, X_m(t_1);\mathbf{p}\right),\\
S_1(t_2) &\approx \dot{X}_1(t_2) = f_1\left(X_1(t_2), X_2(t_2), \ldots, X_m(t_2);\mathbf{p}\right),\\
&\vdots\\
S_1(t_N) &\approx \dot{X}_1(t_N) = f_1\left(X_1(t_N), X_2(t_N), \ldots, X_m(t_N);\mathbf{p}\right),\\
&\vdots\\
S_i(t_k) &\approx \dot{X}_i(t_k) = f_i\left(X_1(t_k), X_2(t_k), \ldots, X_m(t_k);\mathbf{p}\right),\\
&\vdots\\
S_m(t_N) &\approx \dot{X}_m(t_N) = f_n\left(X_1(t_N), X_2(t_N), \ldots, X_m(t_N);\mathbf{p}\right).
\end{aligned} \tag{3.1}$$

Thus, assuming that time-series concentration data of all metabolites $X_i(t_k)$ are available, the slopes $S_i(t_k)$ can be calculated and the estimation simplifies to solving a set of $m \times N$ (nonlinear) algebraic equations, where m is the number of metabolites and N is the number of time points. Note that since there is no integration of the ODEs, the minimization of the difference between slopes and $f(\mathbf{X},\mathbf{p})$ is computationally efficient, even for a large number of parameters. However, one drawback of this method is that the molar balance is only satisfied at discrete time points t_k and thus, the resulting parameter estimates often give concentration time profiles that offset the data.

When data are noisy, slope estimates by finite differencing will have spurious fluctuations as noise is amplified by such calculations. Thus, data smoothing is a necessary step in this method, for example using polynomial fitting, neural network [60] or automated smoother [61]. Regardless of the smoothing methods used, extra care has to be taken to avoid data over-fitting, and even with automated methods, user judgment is still needed in this process.

3.2.2 ODE Decomposition Method

A different decoupling strategy has been proposed that involves solving each of the ODEs one-by-one, and parameter estimates are obtained by minimizing the sum of squares of time-concentration difference between model simulations and data [93,99,101]. During the integration of each ODE, the other states (metabolites) are treated as external inputs, whose values are interpolated from smoothened time-series data. By solving and fitting one metabolite at a time, this method avoids the integration of coupled ODEs and also reduces the parameter search space. In contrast to the decoupling method above, the molar balance of each metabolite is approximately satisfied over time, not just at discrete time points. Furthermore, this method can still be applied in the situation where there are missing metabolite concentrations. However, the ODE stiffness problem, though greatly lessened, is not completely eliminated.

3.2.3 Combined Iterative Estimation

The proposed parameter estimation in this chapter iterates between the aforementioned two methods according to the flowchart shown in Figure 3.1. By doing so, this proposed method combines the computational efficiency of the decoupling method and the reduced search space of the ODE decomposition method, and is also able to handle missing metabolite measurements.

In consideration of missing data of some metabolites, the ODE model is rewritten as:

$$\begin{cases} \dot{\mathbf{X}}_m = f_m\left(\mathbf{X}_m, \mathbf{X}_u; \mathbf{p}_m\right) \\ \dot{\mathbf{X}}_u = f_u\left(\mathbf{X}_m, \mathbf{X}_u; \mathbf{p}_m, \mathbf{p}_u\right) \end{cases}, \tag{3.2}$$

where $\mathbf{X}_m$ and $\mathbf{X}_u$ denote the measured and unmeasured metabolites, respectively, $\mathbf{p}_m$ includes all parameters appearing in f_m, and $\mathbf{p}_u$ includes the remaining parameters (specific to f_u) and the initial concentrations for $\mathbf{X}_u$. Prior to the iteration, data smoothing was performed to reduce noise effects and to obtain slope estimates. Using the smoothened data and initial guesses of the parameters $\mathbf{p}_u$ and $\mathbf{p}_m$, a simulation of unmeasured metabolites is carried out by solving the ODEs for $\mathbf{X}_u$ only, as done in the ODE decomposition method.

The first iteration then begins with the decoupling method to obtain $\mathbf{p}_m$ by minimizing the following slope errors:

$$\sum_{k=1}^{N}\left[\mathbf{S}_m(t_k) - f_m\left(\mathbf{X}_m(t_k), \mathbf{X}_u(t_k); \mathbf{p}_m\right)\right]^T \left[\mathbf{S}_m(t_k) - f_m\left(\mathbf{X}_m(t_k), \mathbf{X}_u(t_k); \mathbf{p}_m\right)\right], \tag{3.3}$$

where $\mathbf{S}_m(t_k)$ is the slope of smoothened data for $\mathbf{X}_m$ at $t=t_k$. Using the estimates of $\mathbf{p}_m$, the values of $\mathbf{p}_u$ are obtained in the next estimation phase by minimizing the concentration errors:

$$\sum_{k=1}^{N}\left[\mathbf{X}_{m,data}\left(t_k\right)-\mathbf{X}_m\left(t_k\right)\right]^T\left[\mathbf{X}_{m,data}\left(t_k\right)-\mathbf{X}_m\left(t_k\right)\right], \tag{3.4}$$

in which all the ODEs are solved one at a time. In this case, the ODEs for $\mathbf{X}_u$ are solved prior to $\mathbf{X}_m$ and the newly simulated $\mathbf{X}_u$ values are then used in the next iteration. If there are more than one unmeasured metabolites, the involved ODEs for $\mathbf{X}_u$ may need to be solved simultaneously. The procedure iterates between the two estimation phases until convergence. Here, the iterations will stop when parameter estimates between iterations differ by less than a chosen convergence factor.

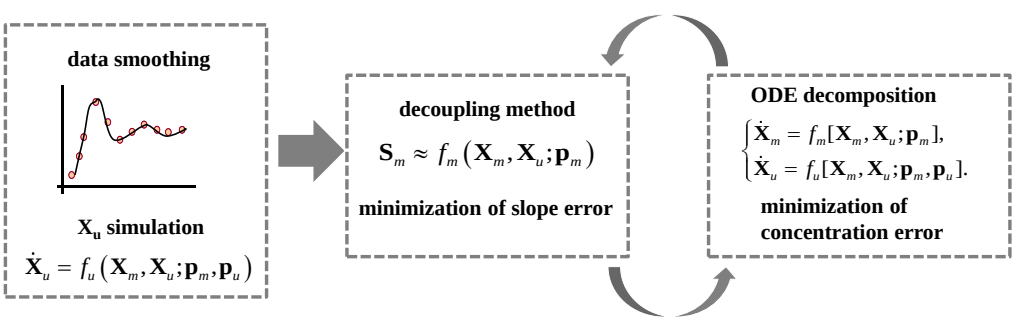


Figure 3.1. Flowchart of the parameter estimation process.

As motivated in Chapter 2, the optimization problems in the two phases are solved using the SSm GO MATLAB toolbox (Scatter Search Method for Global Optimization) [157,158]. In addition, to alleviate the ODE stiffness problem, each

ODE simulation is limited to a given maximum time and those exceeding this upper bound are assigned with a large objective function value.

3.3 Results

The performance of the proposed method is demonstrated in the applications to a generic branched pathway [60], *E. coli* metabolism [70].and the glycolytic pathway of *Lactococcus lactis* (*L. lactis*) [160].

3.3.1 A Generic Branched Pathway

The metabolic pathway in this example is given in Figure 3.2, which describes the transformations among four metabolites (double-line arrows) with feedback activation and inhibition (dashed arrows with plus and minus signs, respectively).

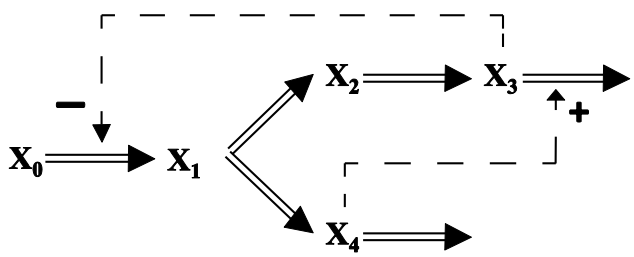


Figure 3.2. A generic branched pathway [60].

This pathway is modeled in the form of an S-system with 12 kinetic parameters, as follows [60]:

$$\begin{cases} \dot{X}_1 = \alpha_1 X_0 X_3^{g_{13}} - \beta_1 X_1^{h_{11}} \\ \dot{X}_2 = \alpha_2 X_1^{h_{11}} - \beta_2 X_2^{h_{22}} \\ \dot{X}_3 = \beta_2 X_2^{h_{22}} - \beta_3 X_3^{h_{33}} X_4^{h_{34}} \\ \dot{X}_4 = \left(\beta_1 - \alpha_2\right) X_1^{h_{11}} - \beta_4 X_4^{h_{44}} \\ X_0 = 0.6 \end{cases} \qquad \begin{matrix} X_1(t_0) = 1.4 \\ X_2(t_0) = 2.7 \\ X_3(t_0) = 1.2 \\ X_4(t_0) = 0.4 \end{matrix} \tag{3.5}$$

This model was used to generate *in silico* noise-free and noisy experimental data (10% additive noise, Gaussian, i.i.d.) using the parameter values reported in the original publication (see Appendix A Table A1) and with the assumption that only X_1, X_2 and X_4 were measured. A *6*-th order polynomial, for which adjusted R^2 reached a maximum, was chosen for data smoothing and to calculate the time-series slopes. The adjusted R^2 was used here to avoid data over-fitting [161]. In the parameter estimation, the search space was limited to α_i, $\beta_i \in [0, 25]$, g_{ij},$h_{ij} \in [-2, 2]$, and $X_3(t_0) \in [0, 5]$. The numerical integrations were performed in MATLAB using ode15s.

One practical issue affecting the parameter estimation in this example lies in that a majority of biological system modeling suffers from the lack of complete parameter identifiability, as discussed in Chapter 2 [64]. In other words, not all parameters can be uniquely identified and only a subset can be determined from data. Here, the proposed method will first be evaluated under the ideal scenario, in which the estimation is done only for the subset of *a priori* identifiable parameters (AIPs) [162] (the other parameters were set to the original values) and using noise-free data. The application of standard least square estimation using fully coupled ODEs encountered numerical stiffness problem and failed to converge. In addition, the decoupling method alone cannot be applied for the estimation involving missing measurements. Thus, in this example, the ODE decomposition estimation was used as a comparison of the proposed method.

Table 3.1 summarizes the estimation results under the ideal scenario described above. In this case, the performance of the proposed method using 0.01%

convergence criterion is comparable to the ODE decomposition alone. The larger parameter deviations in the two-phase estimation are caused by the polynomial smoothing to obtain the time-slope data, without which the performance of the two estimation methods is virtually identical. In addition, by increasing the convergence factor, the proposed method can reduce computational time, but at the cost of increased errors in the parameter estimates.

Table 3.1. Estimation of AIPs in branched pathway model

	ODE Decomposition	**Two-Phase Estimation**		
		0.01% [a]	**0.1%** [a]	**1%** [a]
Computational time (sec) [c]	3968.33	3741.86 (3823.77) [b]	2042.40	810.37
Number of stiff ODE simulations	203	0	0	0
Average parameter error	0.24%	1.37% (0.20%) [b]	2.26%	8.11%
Slope error [d]	2.6435	2.4374 (0.0087) [b]	2.6052	6.6401
Concentration error [e]	0.0047	0.0187 (0.0049) [b]	0.0198	0.0805

[a] Convergence criterion between two estimation phases.

[b] Using slope values evaluated directly using the ODEs.

[c] The computational time was based on Dual Processors Intel Quad-Core 2.83 GHz.

[d] Slope error was calculated using Equation 3.3, in which $\mathbf{X}_u$, $\mathbf{X}_m$ are from simultaneous ODE simulation.

[e] Concentration error was calculated using Equation 3.4, in which $\mathbf{X}_m$ are from simultaneous ODE simulation.

The results of estimating the full parameter set are given in Table 3.2, Figure 3.3 A–B and Appendix A Table A1. Even when data are noise-free, the relative errors of the parameter estimates can reach higher than 300% using the ODE decomposition method. While parameter identifiability issue certainly contributes to these errors, the ODE decomposition in this case failed to extract the maximum information available in the data. As a comparison here, the application of the proposed iterative method using noise-free data and 1% convergence criterion produced improved parameter estimates and importantly, in much shorter time than the ODE decomposition (see Figure 3.4 A–B and Table 3.2). The maximum relative error dropped to 150% and fewer parameters had errors above 50%. In addition, the predicted concentration and slope profiles were relatively better than those from the ODE decomposition alone. While the lack of fit to the missing X_3 data in both methods was expected, parameter estimates from both methods were able to capture the trend of metabolite dynamics.

When using noisy data, the proposed iterative method again gave comparatively more accurate parameter estimates and finished in much shorter time than the ODE decomposition. The results from the two estimation methods are shown in Figures 3.3 C–D and 3.4 C–D and Table 3.2. As expected, these parameter estimates were on average less accurate than those obtained from noise-free data, and the estimation in this case took two to three times longer than those using noise-free data.

Table 3.2. Parameter estimation of the branched pathway model

	ODE Decomposition		Two-Phase Estimation	
	w/o noise	**w/ noise**	**w/o noise**	**w/ noise**
Computational time (sec)	4493.2	10910.3	1062.1	2807.4
Number of stiff ODE simulations	1247	2012	359	823
Average parameter error	92.18%	90.97%	36.59%	47.27%
Slope error	2.5962	9.4303	0.8620	8.5909
Concentration error	0.5137	5.8207	0.1526	3.6021

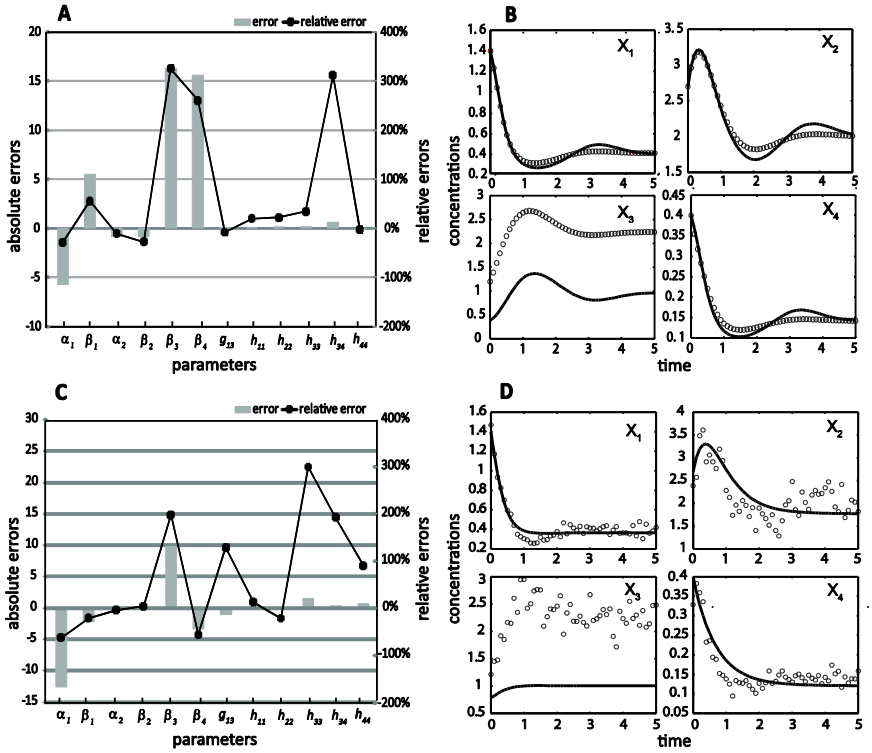


Figure 3.3. ODE decomposition estimation in the branched pathway model: parameter errors (A, C) and concentration simulations (B, D) using noise-free (A, B) and noisy data (C, D); (—) simulation profile, (○) *in silico* data.

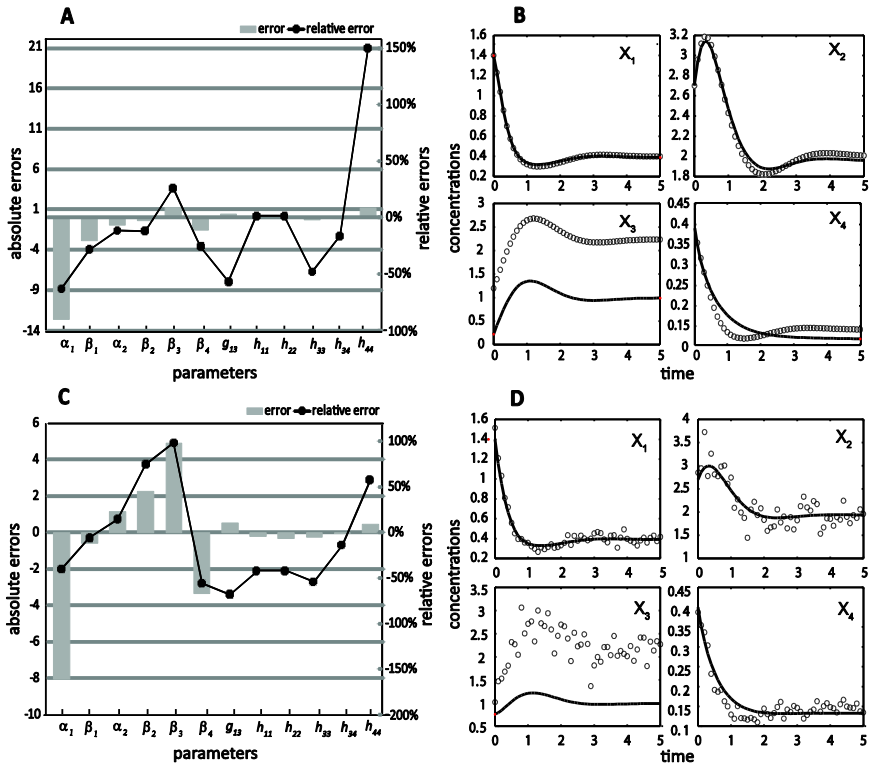


Figure 3.4. Two-phase iterative estimation in the branched pathway model: parameter errors (A, C) and concentration simulations (B, D) using noise-free (A, B) and noisy data (C, D); (—) simulation profile, (○) *in silico* data.

3.3.2 *E. coli* Metabolism Model

The second case study involves a simplified kinetic model of *E. coli* metabolism under glucose feeding [70]. Experimental time-course data (two repeats) were previously reported for two initial glucose concentrations of 40 and 50 g/L (see Figure 3.5). An S-system model was proposed [70] and is given by:

$$\begin{cases} \dot{X}_1 = \alpha_1 X_1^{g_{11}} X_2^{g_{12}} \\ \dot{X}_2 = -\beta_2 X_1^{h_{21}} X_2^{h_{22}} \\ \dot{X}_3 = \alpha_3 X_1^{g_{31}} X_2^{g_{32}} \\ \dot{X}_4 = \alpha_4 X_1^{g_{41}} X_2^{g_{42}} \\ \dot{X}_5 = \alpha_5 X_1^{g_{51}} X_2^{g_{52}} \end{cases} \tag{3.6}$$

where X_1, X_2, X_3, X_4 and X_5 are the concentrations of cell mass, glucose, protein, lactate and acetate, respectively. The four sets of initial concentrations in the experimental data are [0.1645, 39.66, 0.04390, 0, 0] g/L, [0.1156, 38.21, 0.03170, 0, 0] g/L, [0.1931, 48.05, 0.04670, 0, 0] g/L and [0.2227, 51.88, 0.04100, 0, 0] g/L.

In this case, a *4*-th order polynomial data smoothing was chosen using the same maximization of the adjusted R^2 criterion, and the smoothened curves were used to calculate the slopes. In addition, the parameter search space was limited to α_i, β_i, g_{ij}, $h_{ij} \in [10^{-3}, 2]$. Using complete experimental data, the model parameters were first estimated using the decoupling method, which will be used for evaluating estimates from incomplete data. The parameter estimates obtained here were comparable with the values reported in the original publication (see Appendix A Table A3) [70].

In the following, the measurements of X_2 was assumed missing. The parameter search space of α_i, β_i, g_{ij}, h_{ij} remained the same as above and the search space for the two missing initial conditions of X_2 were bounded within [0, 100] g/L. Like in the previous example, for comparison purpose, the ODE decomposition method was applied to obtain parameter estimates from

incomplete data. The results are summarized in Table 3.3 and Figure 3.5. In this example, one can see the severe consequences of sparse experimental data, leading to large slope and concentration errors.

The application of the proposed iterative method again gave improved parameter estimates in shorter amount of optimization time (see Figure 3.6 and Table 3.3). In comparison with the results from the ODE decomposition, the slope and concentration errors were much reduced, at roughly 2.5 times lower computational cost. In addition, the parameter estimates were comparably more accurate, especially for the estimates of rate constants. The average error of parameter estimates was 74.14% for the two-phase method, while the ODE decomposition gave an average error of 119.45 %.

Table 3.3. Parameter estimation of the *E. coli* model

	ODE Decomposition	**Two-Phase Estimation**
Computational time (sec)	66119.7	26855.0
Number of stiff ODE simulations	645	0
Slope error	1863.8	225.82
Concentration error	38083	21809

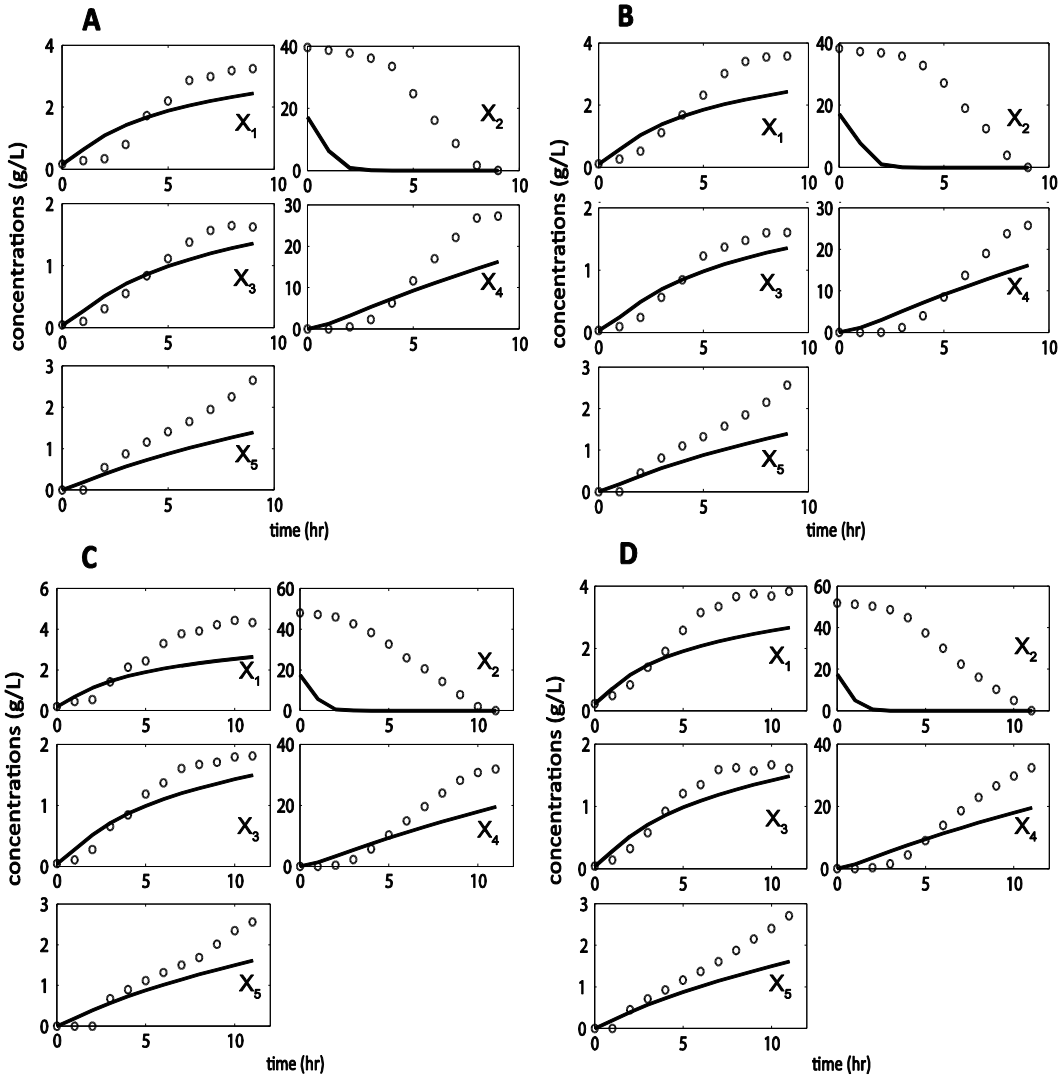


Figure 3.5. ODE decomposition estimation in the *E. coli* model: A and B show concentration simulations of the duplicates with the initial glucose concentration of 40 g/L; C and D show concentration simulations of the duplicates with the initial glucose concentration of 50 g/L. (—) simulation profile, (○) experimental data.

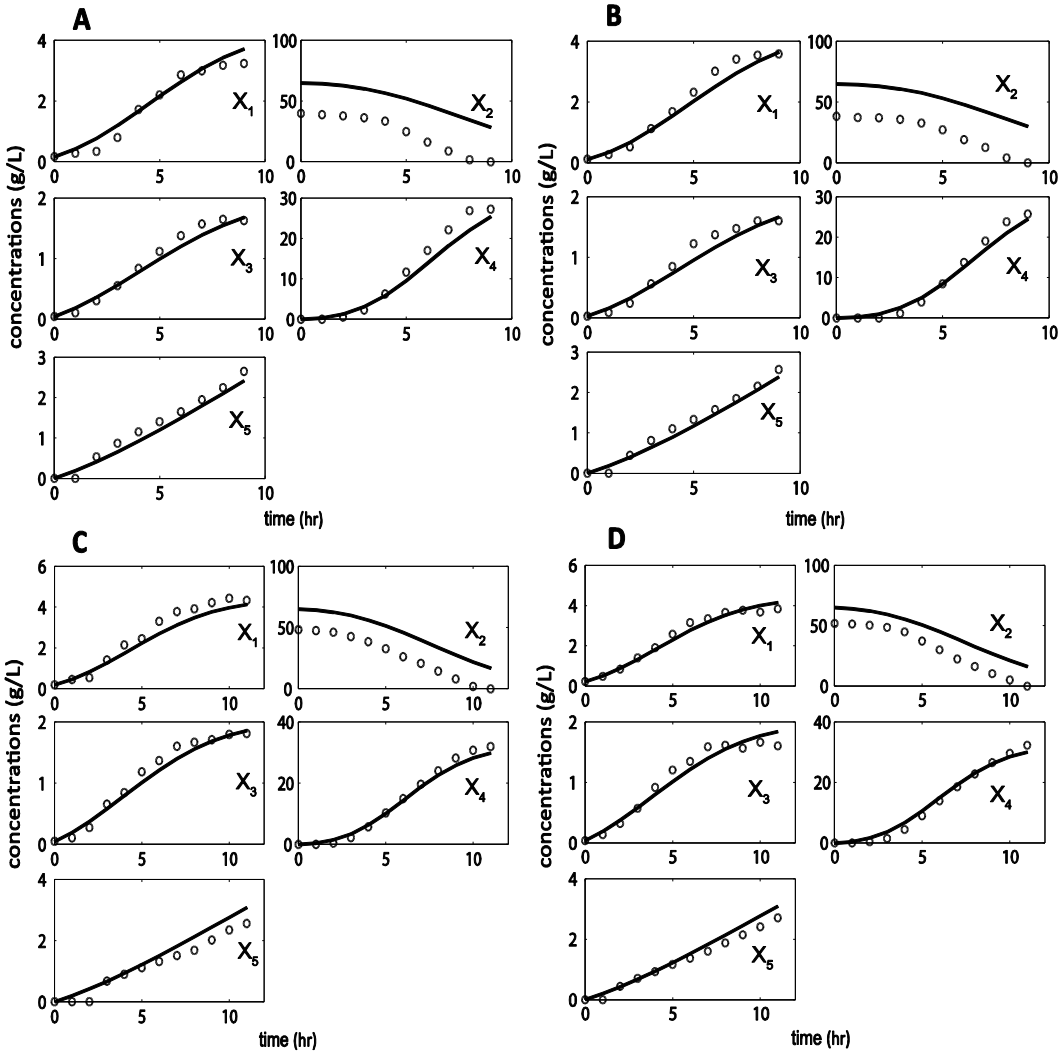


Figure 3.6. Two-phase iterative estimation in the *E. coli* model: A and B show concentration simulations of the duplicates with the initial glucose concentration of 40 g/L; C and D show concentration simulations of the duplicates with the initial glucose concentration of 50 g/L. (—) simulation profile, (○) experimental data.

3.3.3 Glycolytic Pathway in *Lactococcus lactis*

The third case study was taken from the modeling of the glycolytic pathway of *L.lactis*, as shown in Figure 3.7, again using S-system formulism [160]. Experimental time-course data of the metabolites were previously obtained using *in vivo* NMR [76,163]. Here, the concentration variables denote the following

metabolites: glucose (Glu)—X_1, glucose 6-phosphate (G6P)—X_2, fructose 1, 6-biphosphate (FBP)—X_3, phosphoenolpyruvate (PEP)—X_4, lactate (Lac)—X_5, and acetate (Ace)—X_6. Assuming that the known network connectivity is correct, the model equations and initial conditions are given by:

$$\begin{cases} \dot{X}_1 = \alpha_1 - \beta_1 X_1^{h_{11}} X_4^{-h_{14}} & X_1(t_0) = 20 \\ \dot{X}_2 = \alpha_2 X_1^{g_{21}} X_4^{g_{24}} - \beta_2 X_2^{h_{22}} & X_2(t_0) = 0.4 \\ \dot{X}_3 = \alpha_3 X_2^{g_{32}} - \beta_3 X_3^{h_{33}} & X_3(t_0) = 0.4 \\ \dot{X}_4 = \alpha_4 X_3^{g_{43}} - \beta_4 X_2^{h_{42}} X_4^{h_{44}} & X_4(t_0) = 8.5 \\ \dot{X}_5 = \alpha_5 X_4^{-g_{54}} - \beta_5 & X_5(t_0) = 0.05 \\ \dot{X}_6 = \alpha_6 X_4^{-g_{64}} & X_6(t_0) = 0.3 \end{cases} \tag{3.7}$$

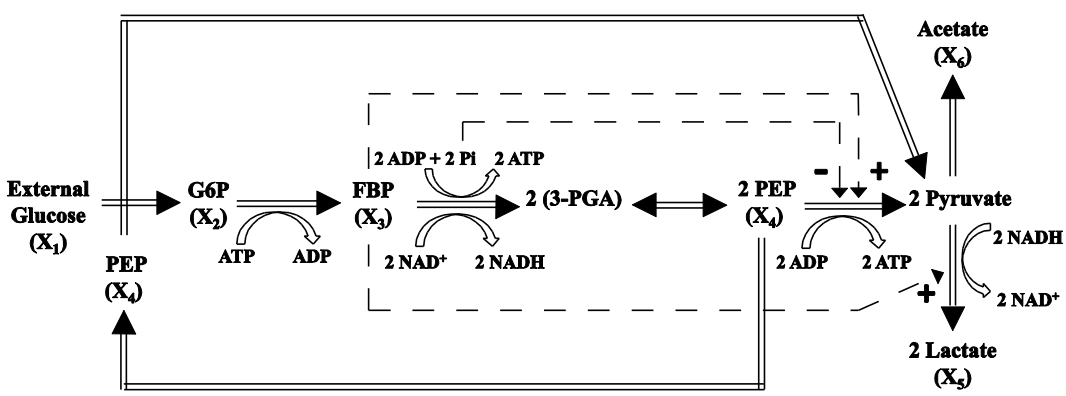


Figure 3.7. The glycolytic pathway in *L. lactis [78]*.

First, using the parameters reported in the original publication [160] (Appendix A Table A4), *in silico* noise-free data were produced for all metabolites except X_3. In this case, we have used a piecewise polynomial fitting, since the data before t=9.4 minute had markedly different dynamics. Specifically,

eighth-order and second-order polynomials were used in the fitting before and after this time, respectively, again based on maximizing the adjusted R^2. The parameter search space was limited such that α_i, $\beta_i \in [0, 20]$, g_{ij}, $h_{ij} \in [0, 5]$ and $X_3(t_0) \in [0, 20]$.

Table 3.4 reports the parameter estimation results using the ODE decomposition and the two-phase iterative method. Compared with the results from the ODE decomposition (Figure 3.8 and Table 3.4), the proposed method gave better concentration and slope fittings at roughly three times lower computational cost. In addition, the average parameter error from the two-phase method was comparably lower. Even with the complete measurements, parameter identifiability issue has been shown to exist in this system [64].

Table 3.4. Parameter estimation of the *L. lactis* metabolic model

	ODE Decomposition		**Two-Phase Estimation**	
	w/o noise	**filtered data**	**w/o noise**	**filtered data**
Computational time (sec)	79772.3	81858.8	24838.9	27325.2
Number of stiff ODE simulations	875	1023	316	368
Average Parameter error	243.90%	—	97.29%	—
Slope error (1/ N [a])	77.350	27090.2	2.3240	1.4910
Concentration error (1/ N [a])	24.777	288.71	24.784	24.573

[a] N is the number of time points in each metabolic profile.

Finally, the two-phase iterative estimation and the ODE decomposition were applied to the published smoothened NMR data using an automated smoother [61]. Again without X_3, the estimation results are summarized in Table 3.4 and illustrated in Figure 3.9. As before, the proposed method gave markedly improved concentration and slope fittings in a shorter amount of time than the ODE decomposition method.

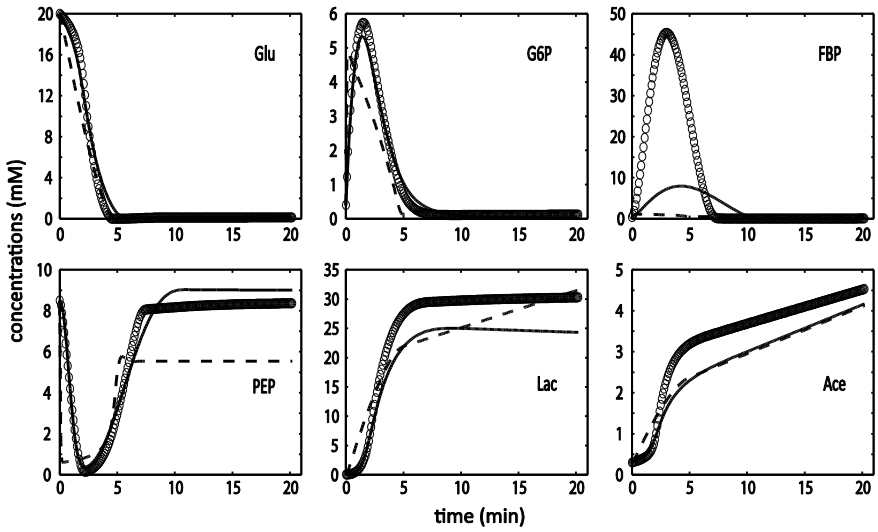


Figure 3.8. Metabolic profiles in the *L.lactis* glycolytic pathway: *in silico* data (open circles), ODE decomposition (dashed line), and two-phase iterative estimation (solid line).

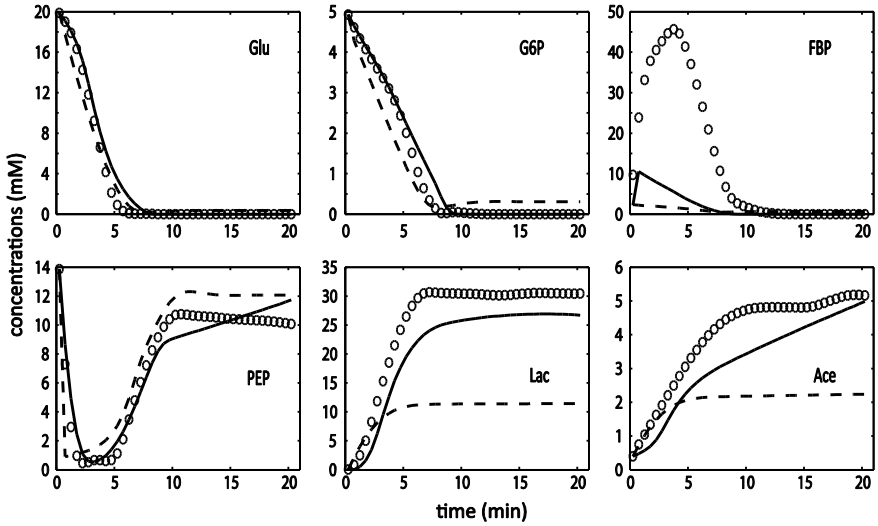


Figure 3.9. Metabolic profiles in the *L. lactis* glycolytic pathway: smoothened data (open circles), ODE decomposition (dashed line), and two-phase iterative estimation (solid line).

3.4 Discussion

The proposed iterative parameter estimation method in this chapter builds on the strengths of the decoupling method and the ODE decomposition method. By decoupling the ODEs, this proposed method is significantly faster than other methods that require integrating the coupled ODEs for each objective function evaluation, while still giving good fit to measured concentration data. In addition, like the ODE decomposition method, the combined approach does not require complete measurements of all metabolites and has much reduced the parameter search space. As shown in the applications to the three cases, the proposed method was superior to the two methods from which it was developed. When metabolite measurements were incompletely available, the decoupling method could not be applied. Compared with the ODE decomposition method, the proposed method gave more accurate parameter estimates and better data fit (slope and concentration) at a much lower computational cost. While the fit to missing concentration measurement unsurprisingly had an offset, it is noteworthy that the dynamic trend can still be captured.

The combination of slope and concentration fittings had also been used in several existing parameter estimation methods. For example, Wang and Liu had developed a method where kinetic parameters were estimated simultaneously by minimizing both slope and concentration errors using a multi-objective optimization framework [164]. Similar to the two-phase method here, Gennemark and Wedelin had proposed a multi-step method, where a derivative method was

used to obtain initial, rough guesses of model parameters and a subsequent minimization of concentration error was performed starting from these guesses [165]. However, these two methods assumed that all metabolite measurements are available. Notably, in the latter, the ODEs were also solved one at a time using single or multiple shooting methods, thereby decoupling the parameter estimation problem as in the ODE decomposition. The shooting method can in fact be used to substitute the role of ODE decomposition method in the two-phase iterative estimation here, providing an alternative method.

Another method extended a class of ODE solvers, called orthogonal collocation method, for estimating model parameters [166]. In this case, the concentrations were approximated as a linear combination of basis functions, where the coefficients were treated as nuisance parameters. Model parameters were then simultaneously estimated by minimizing the approximation errors between the simulated concentrations and the data, and between the time-derivatives of concentration predictions and the right hand side of $\dot{\mathbf{X}}(t) = f\left(\mathbf{X}(t);\mathbf{p}\right)$. Despite the similarities, the proposed method differs from this and the aforementioned methods in the grouping parameters into two, those associated with measured variables and those with unmeasured concentrations. By doing so, the parameter estimation can be achieved more efficiently. This is because solving a few small parameter estimation problems is easier than solving the simultaneous estimation of the combined parameter set. In addition, if more metabolites are measured, the estimation naturally becomes faster, since more parameters will be estimated in the first computationally efficient phase.

Although the proposed method performed better than the ODE decomposition method in terms of data fitting (i.e., smaller slope and concentration errors), many of the parameter estimates were still far from the true values (see Appendix A, Tables A1, A3 and A4). This may not be surprising as that the estimation problems had assumed missing data for metabolites. Nevertheless, even with complete data, parameter identifiability has been shown to be lacking in the estimation of kinetic parameters from time-series data and the severity of this problem can be assessed quantitatively [63,64].

Related to the identifiability issue, the kinetic information contained in different metabolites are not equal. The expected degradation in the accuracy of the parameter estimates from missing data depends on the degree of connectivity of the missing metabolite(s) in two ways. The kinetic information (i.e., rate of change) of a metabolite is partially contained in the downstream and upstream metabolites in the metabolic network. The higher the degree of connectivity, by stoichiometry, of a missing metabolite, the more can the missing flux information be re-extracted from the available data. While this missing flux can be determined, the (initial) concentration of the unmeasured metabolite however is still lost. Thus, it is possible to capture the trend of the missing profiles, but not the absolute concentration values, giving an offset between the simulated and true concentrations, as seen in the first and second and to some degree in the third example above.

However, when considering regulatory connectivity, the concentration of metabolite(s) is important. Here, the loss of concentration data of an important,

highly connected regulatory metabolite will lead to a significant loss of information that cannot be easily recovered. In the first example, the loss of metabolite X_3 data represented the worst-case scenario, as this metabolite has a high regulatory connectivity and missing downstream metabolite data. On the other hand, if X_2 was not measured, the parameters can still be identified from other metabolites, since the set of $\mathbf{p}_u$ is null, i.e., the estimation can be done using only the decoupling method. Finally, an increase in the number of unmeasured metabolites will, in general, lead to lower overall kinetic information and poorer parameter estimates. In the first example, missing both X_2 and X_3 indeed gave less accurate parameter estimates, but the two-phase method still outperformed the ODE decomposition (see Appendix A, Tables A1, A2 and Figure A1).

For a given system, the computational requirement of the proposed method depends on several aspects, such as the numbers of measured and unmeasured metabolites, the numbers of parameters associated with measured and unmeasured metabolites, the convergence speed of the iterations, and as seen in the examples, the noise in the data. In general, the higher the number of parameters involved in the first phase, the faster will the estimation finish. Unfortunately, the scalability of the method to larger systems is difficult to be determined, as all of the factors mentioned above will interact. For example, the scaling will depend on the distribution of the additional parameters between the two phases as well as on the dynamics of the system (e.g., related to stiffness of the ODEs). In addition, the convergence will also play an important role, but unfortunately, this is difficult to consider as the two phases have different objective functions.

Finally, while the applications considered in this chapter were taken from S-system models, the proposed iterative estimation is not limited to power-law models. The reason to consider these examples was that these models represent some of the most difficult parameter estimation problems due to the large number of parameters, stiff ODEs and high degree of nonlinearity. The proposed method can also be applied to the problems in which complete time-series data are available. In such cases, the parameters can be divided into two groups based on the level of difficulty in estimating them in each estimation phase. For example, for S-system models, the kinetic orders can be grouped together in the first phase (decoupling method), while the rate constants can be estimated in the second phase (ODE decomposition).

CHAPTER 4 : INCREMENTAL PARAMETER ESTIMATION OF KINETIC METBABOLIC NETWORK MODELS

4.1 Summary

Most of the existing parameter estimation methods involve finding the global minimum of data fitting residuals over the entire parameter space simultaneously. As discussed in Chapter 2, the associated computational requirement often becomes prohibitively high due to the large number of parameters and the lack of complete parameter identifiability (i.e. not all parameters can be uniquely identified).

In this chapter, an incremental approach is applied to the parameter estimation of ODE models from time-concentration profiles. Particularly, the method is developed to address a commonly encountered circumstance in the modeling of metabolic networks, where the number of metabolic fluxes (reaction rates) exceeds that of metabolites (chemical species). Here, the minimization of model residuals is performed over a subset of the parameter space that is associated with the degrees of freedom (DOFs) in the dynamic flux estimation from the concentration time-slopes. The efficacy of this method was demonstrated using two generalized mass action (GMA) models, where the method significantly outperformed single-step estimations. In addition, an extension of the estimation method to handle missing data is also presented. The proposed incremental

estimation method is able to tackle the issue on the lack of complete parameter identifiability and to significantly reduce the computational efforts in estimating model parameters, which will facilitate kinetic modeling of genome-scale cellular metabolism in the future.

4.2 Method

The GMA model of cellular metabolism describes the mass balance of metabolites, taking into account all metabolic influxes and effluxes, as given in Equation 1.1 and rewritten here for reference:

$$d\mathbf{X}(t,\mathbf{p})/dt = \dot{\mathbf{X}}(t,\mathbf{p}) = \mathbf{S}\mathbf{v}(\mathbf{X},\mathbf{p}), \tag{4.1}$$

where $\mathbf{X}(t,\mathbf{p})$ is the vector of metabolic time-concentration profiles, $\mathbf{S} \in \mathbf{R}^{m\times n}$ is the stoichiometric matrix for m metabolites that participate in n reactions, and $\mathbf{v}(\mathbf{X},\mathbf{p})$ denotes the vector of metabolic fluxes (i.e., reaction rates). As introduced in Chapter 1, according to Biochemical Systems Theory, each flux is described by a power-law equation:

$$v_j(\mathbf{X},\mathbf{p}) = \gamma_j \prod_i X_i^{f_{ji}}, \tag{4.2}$$

where γ_j is the rate constant of the j-th flux and f_{ji} is the kinetic order parameter, representing the influence of metabolite X_i on the j-th flux (positive: activation or substrate, negative: inhibition).

In the incremental parameter estimation, noisy time-course concentration data $\mathbf{X}_m(t_k)$ are usually smoothened before the approximation of time-slopes $\dot{\mathbf{X}}_m(t_k)$. Subsequently, the dynamic metabolic fluxes $\mathbf{v}(t_k)$ are estimated from Equation 4.1 by substituting $\dot{\mathbf{X}}(t)$ with $\dot{\mathbf{X}}_m(t_k)$. Finally, the kinetic parameters associated with the j-th flux (i.e., γ_j and f_{ji}'s) can be calculated using a least square regression of

the power-law flux function in Equation 4.2 against the estimated $v_j(t_k)$. Note that for GMA models, the least square parameter regressions in the last step are essentially linear in the logarithmic scale and thus, can be performed very efficiently.

A unique set of dynamic flux values $\mathbf{v}(t_k)$ can only be computed from $\dot{\mathbf{X}}_m(t_k) = \mathbf{S}\mathbf{v}(t_k)$, when the number of metabolites exceeds (or equals) that of fluxes. However, a metabolite in general can participate in more than one metabolic flux ($m < n$). In such a situation, there exist an infinite number of dynamic flux combinations $\mathbf{v}(t_k)$ that satisfy $\dot{\mathbf{X}}_m(t_k) = \mathbf{S}\mathbf{v}(t_k)$. The dimensionality of the set of the flux solutions is equal to the DOF, given by the difference between the number of fluxes and the number of metabolites: $n_{DOF} = n - m > 0$ (assuming **S** has a full row rank, i.e., there is no redundant ODE in Equation 4.1). The positive DOF means that the values of n_{DOF} selected fluxes can be independently set, from which the remaining fluxes can be computed. This relationship forms the basis of the proposed estimation method, in which the model goodness of fit to data is optimized by adjusting only a subset of parameters associated with the independent fluxes.

Specifically, it is started by decomposing the fluxes into two groups: $\mathbf{v}(t_k) = [\ \mathbf{v}_I(t_k)^T \quad \mathbf{v}_D(t_k)^T\ ]^T$, where the subscripts I and D denote the independent and dependent subsets, respectively. Then, the parameter vector **p** and the stoichiometric matrix **S** can be structured correspondingly as $\mathbf{p} = [\ \mathbf{p}_I \ \ \mathbf{p}_D\ ]$ and **S** =

[$\mathbf{S}_I$ $\mathbf{S}_D$]. The relationship between the independent and dependent fluxes can be formulated by rearranging $\dot{\mathbf{X}}_m(t_k) = \mathbf{S}\mathbf{v}(t_k)$ into:

$$\mathbf{v}_D(t_k) = \mathbf{S}_D^{-1}\left[\dot{\mathbf{X}}_m(t_k) - \mathbf{S}_I\mathbf{v}_I\left(\mathbf{X}_m(t_k), \mathbf{p}_I\right)\right]. \tag{4.3}$$

In this case, given $\mathbf{p}_I$, one can compute the independent fluxes $\mathbf{v}_I(\mathbf{X}_m(t_k),\mathbf{p}_I)$ using the concentration data $\mathbf{X}_m(t_k)$, and subsequently obtain $\mathbf{v}_D(t_k)$ from Equation 4.3. Finally, $\mathbf{p}_D$ can be estimated by a simple least square fitting of $\mathbf{v}_D(\mathbf{X}_m(t_k),\mathbf{p}_D)$ to the computed $\mathbf{v}_D(t_k)$ one flux at a time, when there are more time points than the number of parameters in each flux.

In this work, two formulations of the parameter estimation of ODE models in Equation 4.1 are investigated, involving the minimization of concentration and slope errors. The objective function for the concentration error is given by

$$\Phi_C(\mathbf{p},\mathbf{X}) = \frac{1}{mK}\sum_{k=1}^{K}\left[\mathbf{X}_m(t_k) - \mathbf{X}(t_k,\mathbf{p})\right]^T\left[\mathbf{X}_m(t_k) - \mathbf{X}(t_k,\mathbf{p})\right] \tag{4.4}$$

and that for the slope error is given by

$$\Phi_S(\mathbf{p},\mathbf{X}) = \frac{1}{mK}\sum_{k=1}^{K}\left[\dot{\mathbf{X}}_m(t_k) - \mathbf{S}\mathbf{v}(\mathbf{X}_m(t_k),\mathbf{p})\right]^T\left[\dot{\mathbf{X}}_m(t_k) - \mathbf{S}\mathbf{v}(\mathbf{X}_m(t_k),\mathbf{p})\right], \tag{4.5}$$

where K denotes the total number of measurement time points and $\mathbf{X}(t_k,\mathbf{p})$ is the concentration prediction (i.e., the solution to the ODE model in Equation 4.1). Figure 4.1 describes the formulation of the incremental parameter estimation and the procedure for computing these objective functions. Note that the computation of Φ_C requires integrations of the ODE model and thus, the estimation using this

objective function is expected to be computationally costlier than that using Φ_S. On the other hand, metabolic mass balance is only approximately satisfied at discrete time points t_k during the parameter estimation using Φ_S, as the ODE model is not integrated.

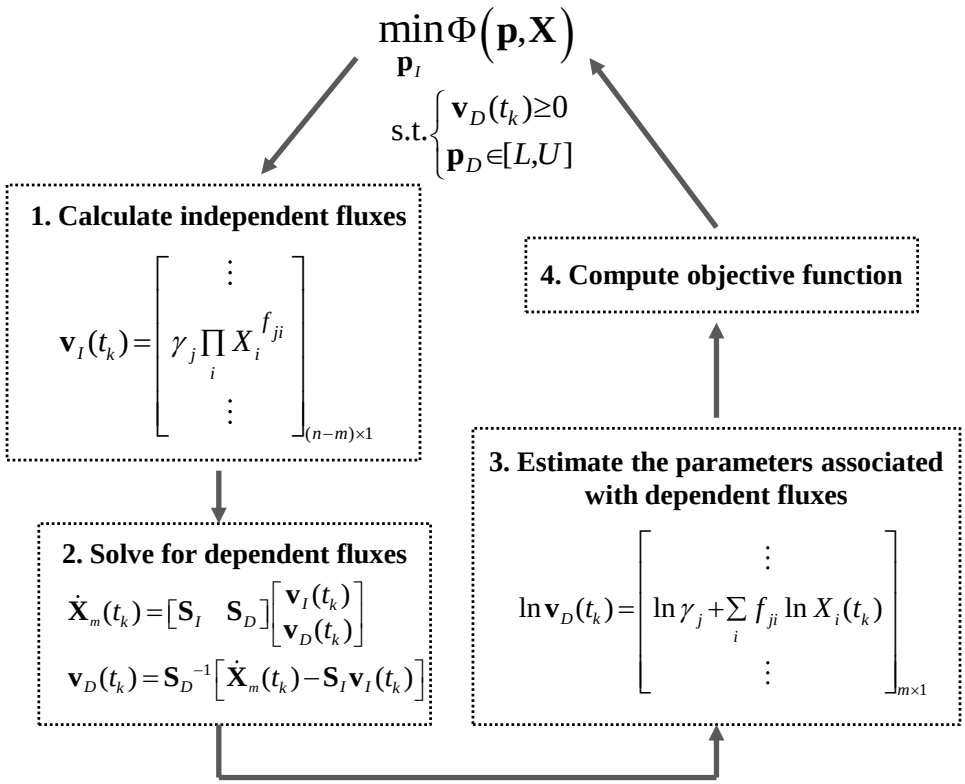


Figure 4.1. Flowchart of the incremental parameter estimation.

There are several important practical considerations in the implementation of the proposed method. The first consideration is on the selection of the independent fluxes. Here, the set of these fluxes is selected such that (i) the $m \times m$ submatrix $\mathbf{S}_D$ is invertible, (ii) the total number of the independent parameters $\mathbf{p}_I$ is small, and (iii) the prior knowledge of the corresponding $\mathbf{p}_I$ is maximized. The

last two aspects should lead to a reduction in the parameter search space as well as the cost of finding the global optimal solution of the minimization problem in Figure 4.1. The second consideration is regarding constraints in the parameter estimation. Biologically relevant values of parameters are often available, providing lower and/or upper bounds for the parameter estimates. In addition, enzymatic reactions in the ODE model are typically assumed to be irreversible and thus, dynamic flux estimates are constrained to be positive. Hence, the parameter estimation involves a constrained minimization problem, for which many global optimization algorithms exist.

So far, it has been assumed that the time-course concentration data are available for all metabolites. However, the method introduced above can be modified to accommodate more general circumstances, in which data for one or several metabolites are missing. Like in Chapter 3, the ODE model is first rewritten to separate the mass balances associated with measured and unmeasured metabolites, such that

$$\dot{\mathbf{X}}(t,\mathbf{p}) = \begin{bmatrix} \dot{\mathbf{X}}_M \\ \dot{\mathbf{X}}_U \end{bmatrix}(t,\mathbf{p}) = \begin{bmatrix} \mathbf{S}_M \\ \mathbf{S}_U \end{bmatrix} \mathbf{v}(\mathbf{X}_M, \mathbf{X}_U, \mathbf{p}) \tag{4.6}$$

where the subscripts M and U refer to components that correspond to measured and unmeasured metabolites, respectively. Again, if the fluxes are split into two categories $\mathbf{v}_I$ and $\mathbf{v}_D$ as above, the following relationship still applies for the measured metabolites:

$$\mathbf{v}_D(t_k) = \mathbf{S}_{D,M}^{-1}\left[\dot{\mathbf{X}}_m(t_k) - \mathbf{S}_{I,M}\mathbf{v}_I(t_k)\right] \tag{4.7}$$

Naturally, the degrees of freedom associated with the dynamic flux estimation is higher by the number of components in $\mathbf{X}_U$ than before. Figure 4.2 presents a modified version of the parameter estimation procedure in Figure 4.1 to handle the case of missing metabolic profiles, in which an additional step involving the simulation of unmeasured metabolites $\dot{\mathbf{X}}_U = \mathbf{S}_U \mathbf{v}\left(\mathbf{X}_M, \mathbf{X}_U, \mathbf{p}\right)$ will be performed. In this integration, $\mathbf{X}_M$ is treated as an external input, whose time-profiles are interpolated from the measured concentrations. The set of independent fluxes $\mathbf{v}_I$ are now selected to include all the fluxes that appear only in $\dot{\mathbf{X}}_U$ and those that lead to a full column ranked $\mathbf{S}_{D,M}$. If $\mathbf{S}_{D,M}$ is a non-square matrix, then a pseudo-inverse will be done in Equation 4.7. Of course, the same considerations discussed above are equally relevant in this case. Note that the initial conditions of $\mathbf{X}_U$ will also need to be estimated.

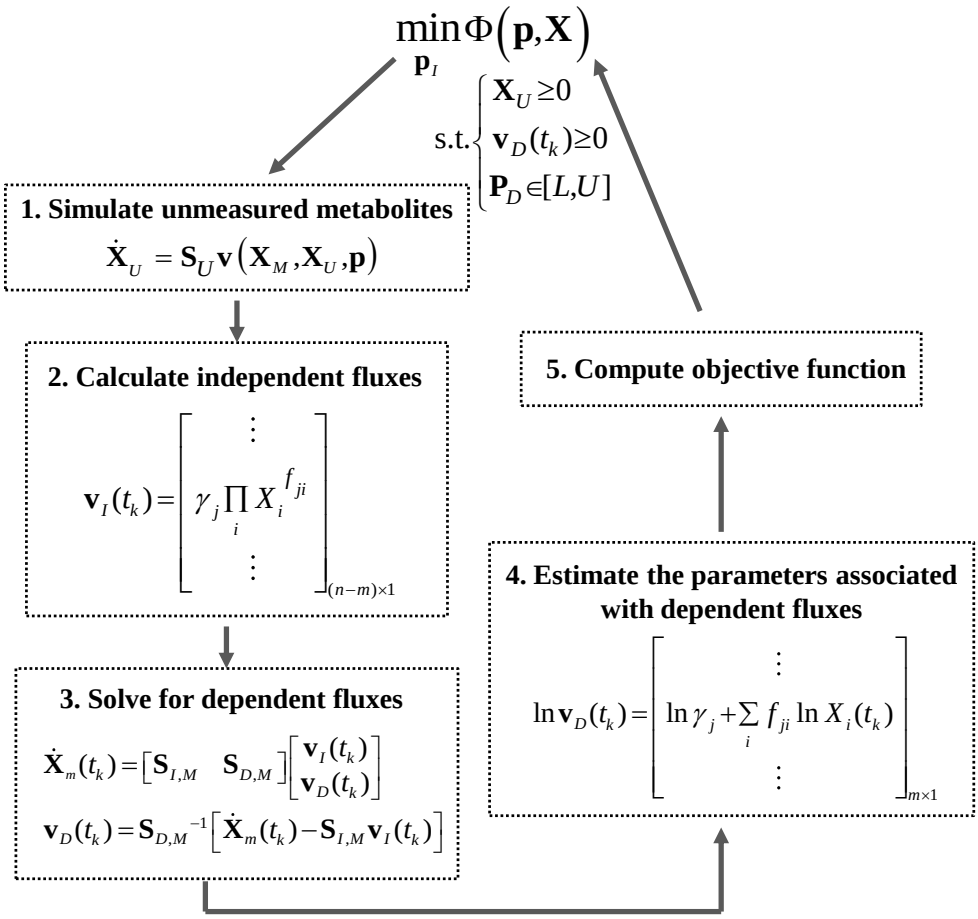


Figure 4.2. Flowchart of the incremental parameter estimation when metabolites are not completely measured.

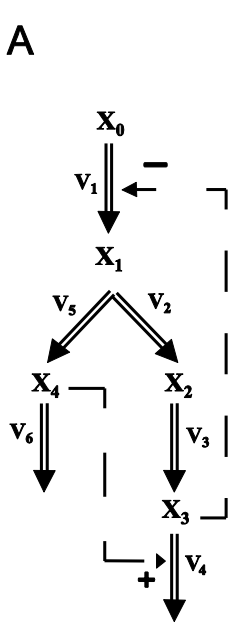
4.3 Results

Two case studies: a generic branched pathway [60] and the glycolytic pathway of *L. lactis* [78], were used to evaluate the performance of the proposed estimation method. In addition, simultaneous estimation methods employing the same objective functions given in Equations 4.4 and 4.5 were applied to these case studies for comparison and to gauge the reduction in the computational cost from using the proposed strategy. In order to alleviate the ODE stiffness issue, parameter combinations that lead to a violation in the MATLAB (ode15s) integration time step criterion is assigned a large error value ($\Phi_C = 10^3$ for the branched pathway and 10^5 for the glycolytic pathway). Alternatively, one could also set a maximum allowable integration time and penalize the associated parameter values upon violation, as described above. In this work, the optimization problems were performed in MATLAB again using the eSSM GO (Enhanced Scatter Search Method for Global Optimization) toolbox [157,158]. Each parameter estimation was repeated five times to ensure the reliability of the global optimal solution. Unless noted differently, the iterations in the optimization algorithm were terminated when the values of objective functions improve by less than 0.01% or the runtime has exceeded the maximum duration (5 days).

4.3.1 A Generic Branched Pathway

The generic branched pathway in this example, which is the same as the one used in the first case study of Chapter 3, consists of four metabolites and six

fluxes, describing the transformations among the metabolites (double-line arrows), with feedback activation and inhibition (dashed arrows with plus or minus signs, respectively), as shown in Figure 4.3 A). The GMA model version of this pathway is given in Figure 4.3 B, containing a total of thirteen rate constants and kinetic orders. This model with the parameter values and initial conditions reported previously [60] were used to generate noise-free and noisy time-course concentration data (i.i.d additive noise from a Gaussian distribution with 10% coefficient of variation). The noisy data were then smoothened using a *6*-th order polynomial, which provided the best relative goodness of fit among polynomials according to Akaike Information Criterion (AIC) [167] and adjusted R^2 [161]. Subsequently, time-slopes of noise-free and smoothened noisy data were computed using central finite difference approximation.



B

$$\begin{bmatrix} \dot{X}_1 \\ \dot{X}_2 \\ \dot{X}_3 \\ \dot{X}_4 \end{bmatrix} = \begin{bmatrix} 1 & -1 & 0 & 0 & -1 & 0 \\ 0 & 1 & -1 & 0 & 0 & 0 \\ 0 & 0 & 1 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & -1 \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \\ v_3 \\ v_4 \\ v_5 \\ v_6 \end{bmatrix}$$

$$v_1 = \gamma_1 X_0 X_3^{-f_{13}}$$
$$v_2 = \gamma_2 X_1^{f_{21}}$$
$$v_3 = \gamma_3 X_2^{f_{32}}$$
$$v_4 = \gamma_4 X_3^{f_{43}} X_4^{f_{44}}$$
$$v_5 = \gamma_5 X_1^{f_{51}}$$
$$v_6 = \gamma_6 X_4^{f_{64}}$$
$$X_0 = 0.6$$

$$X_1(t_0) = 1.40$$
$$X_2(t_0) = 2.70$$
$$X_3(t_0) = 1.20$$
$$X_4(t_0) = 0.40$$

Figure 4.3. A generic branched pathway: (A) Metabolic pathway map and (B) The GMA model equations.

Here, v_1 and v_6 were chosen as the independent fluxes as they comprise the least number of kinetic parameters and lead to an invertible $\mathbf{S}_D$. The two rate constants and two kinetic orders were constrained to within [0, 25] and [0, 2], respectively. In addition, all the reactions are assumed to be irreversible.

Table 4.1 compares simultaneous and incremental parameter estimation runs using noise-free data, employing the two objective functions introduced above. Regardless of the objective functions, the proposed incremental approach significantly outperformed the simultaneous estimation. When using the concentration-error minimization, simultaneous optimization met great difficulty to converge due to stiff ODE integrations, as discussed in Chapter 2. Only one out of five repeated runs could complete after relaxing the convergence criterion of the objective function to 1%, while the others were prematurely terminated after the prescribed maximum runtime of 5 days. In contrast, the proposed incremental estimation was able to find a minimum of Φ_C in less than 120 seconds with reasonably good concentration fit and parameter accuracy (see Figure 4.4 A and Table 4.1). By avoiding ODE integrations and minimizing Φ_S instead, the simultaneous estimation of parameters could be completed in roughly 10 minutes duration, but this was still slower than the incremental estimation using Φ_C. In this case, the incremental method was able to converge in under 2 seconds or over 250 times faster than the simultaneous estimation counterpart was. The goodness of fit to concentration data and the accuracy of parameter estimates were relatively equal for all three completed estimations (see Figure 4.4 B and Table 4.1). The parameter inaccuracy in this case was mainly due to the polynomial smoothing of

the concentration data, since the same estimations using the analytical values of the slopes (by evaluating the right hand side of the ODE model in Equation 4.1) could give accurate parameter estimates (see Appendix B Table B1).

Table 4.1. Parameter estimations of the branched pathway model using noise-free data

	Simultaneous method		**Incremental method**	
	$\min \Phi_C$ [b]	$\min \Phi_S$ [c]	$\min \Phi_C$ [c]	$\min \Phi_S$ [c]
Computational time (sec) [a]	56.00 h	620.81 ± 64.30	95.95 ± 11.09	1.56 ± 0.19
Average parameter error	49.10%	36.91% ± 1.09%	21.56% ± 7.57×10^{-2}%	36.85% ± 6.48×10^{-3}%
$\sqrt{\Phi_C}$ [d]	$\underline{4.54\times10^{-3}}$	6.54×10^{-3} ± 5.20×10^{-5}	$\underline{4.03\times10^{-3}}$ ± 6.22×10^{-8}	6.00×10^{-3} ± 5.05×10^{-7}
$\sqrt{\Phi_S}$ [d]	7.01×10^{-2}	$\underline{2.72\times10^{-2}}$ ± 1.09×10^{-5}	3.92×10^{-2} ± 9.86×10^{-6}	$\underline{2.76\times10^{-2}}$ ± 4.46×10^{-10}

a. The computational time was based on a workstation with dual Intel Quad-Core 2.83 GHz processors.

b. Only one out of five runs completed with a relative improvement of the objective function below 1% between iterations. The rest did not converge within the 5-day time limit after iterating for 583, 989, 777, and 661 times. The corresponding Φ_C at termination were 4.85×10^{-2}, 1.39×10^{-2}, 1.75×10^{-2} and 3.75×10^{-2}, respectively.

c. Mean value and standard deviation (±) out of five repeats, which converged with relative improvement of the objective function below 0.01%.

d. Root mean square error of model predictions, where the underlined value refers to the objective function of the minimization.

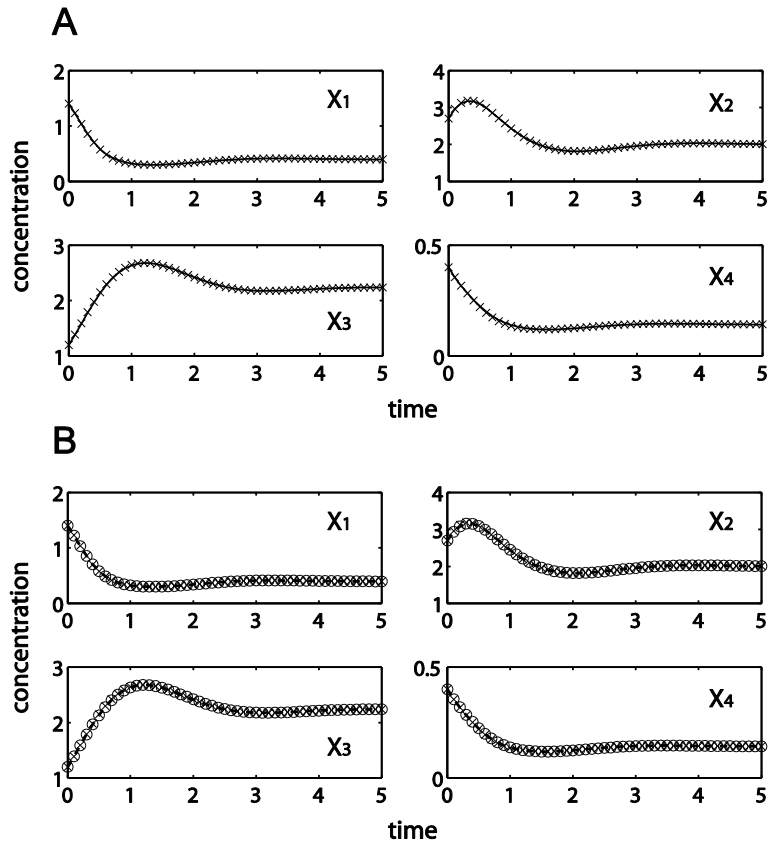


Figure 4.4. Simultaneous and incremental estimation of the branched pathway using *in silico* noise-free data (×). (A) Concentration predictions using parameter estimates from incremental method by Φ_C minimization (—); (B) Concentration predictions using parameter estimates from simultaneous method (○) and proposed method (- - -) by Φ_S minimization.

Table 4.2 provides the results of the same estimation procedures as above using noisy data. Data noise led to a loss of information and an expected decline in the parameter accuracy. Like before, the simultaneous estimation using Φ_C met stiffness problem and three out of five runs did not finish within the five-day time limit. The incremental approach using either one of the objective functions offered a significant reduction in the computational time over the simultaneous estimation using Φ_S, while providing comparable parameter accuracy and concentration and slope fittings (see Figure 4.5 and Table 4.2). In this example,

data noise did not affect the computational cost in obtaining the (global) minimum of the objective functions.

Table 4.2. Parameter estimations of the branched pathway model using noisy data.

	Simultaneous method		**Incremental method**	
	$\min \Phi_C$ [a]	$\min \Phi_S$	$\min \Phi_C$	$\min \Phi_S$
Computational time (sec)	17.86 h 44.63 h	534.83 ± 22.12	71.88 ± 6.33	1.17 ± 0.12
Average parameter error	75.42% 34.98%	54.36% ± 4.47%	75.77% ± 6.11×10^{-3}%	51.15% ± 1.38×10^{-3}%
$\sqrt{\Phi_C}$	$\underline{3.62\times10^{-2}}$ $\underline{3.27\times10^{-2}}$	6.06×10^{-2} ± 1.14×10^{-3}	$\underline{3.52\times10^{-2}}$ ± 9.50×10^{-9}	4.76×10^{-2} ± 3.81×10^{-7}
$\sqrt{\Phi_S}$	2.06×10^{-1} 1.60×10^{-1}	$\underline{1.34\times10^{-1}}$ ± 6.02×10^{-4}	1.64×10^{-1} ± 2.23×10^{-5}	$\underline{1.38\times10^{-1}}$ ± 2.26×10^{-10}

a. Two out of five runs completed with a relative improvement of the objective function below 1% between iterations. The rest did not converge within the 5-day time limit after iterating for 805, 699, and 568 times. The corresponding Φ_C at termination were 4.08×10^{-2}, 5.05×10^{-2} and 6.25×10^{-2}, respectively.

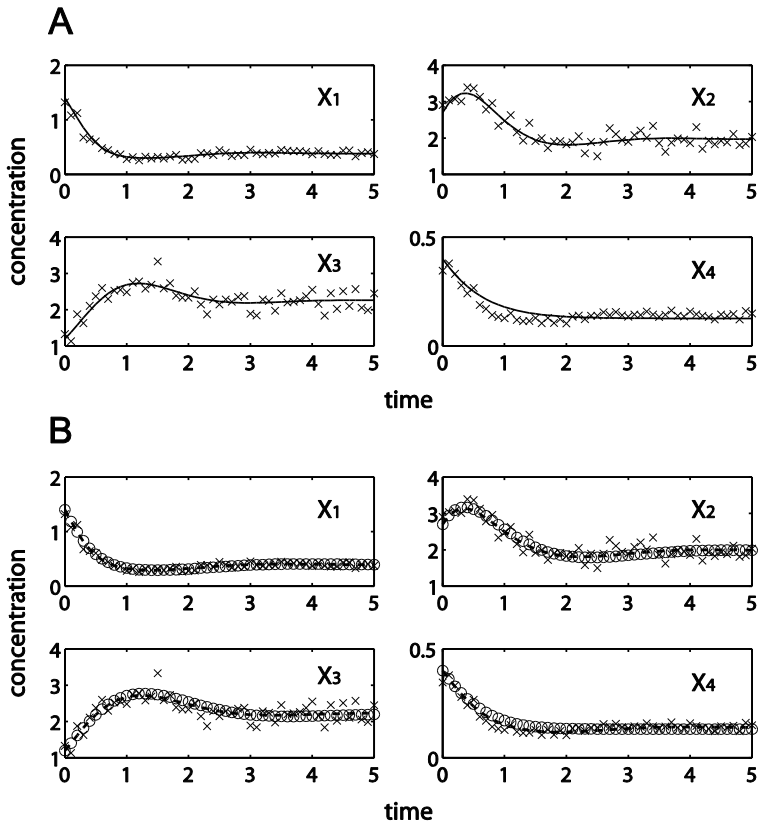


Figure 4.5. Simultaneous and incremental estimation of the branched pathway using *in silico* noisy data (×). (A) Concentration predictions using parameter estimates from incremental method by Φ_C minimization (—); (B) Concentration predictions using parameter estimates from simultaneous method (○) and proposed method (- - -) by Φ_S minimization.

Finally, the estimation strategy described in Figure 4.2 was applied to this example using noise-free data and assuming X_3 data were missing. Fluxes v_3 and v_4 that appear in $\dot{X}_3$ were chosen to be among the independent fluxes and flux v_1 was also added to the set such that the dependent fluxes can be uniquely determined from Equation 4.7. In addition to the parameters associated with the aforementioned fluxes, the initial condition $X_3(t_0)$ was also estimated. The bounds for the rate constants and kinetic orders were kept the same as above, while the initial concentration was bounded within [0, 5].

Table 4.3 summarizes the parameter estimation results. Four out of five repeated runs of Φ_C simultaneous optimization were prematurely terminated after 5 days. Meanwhile, the rest of the estimations could provide reasonably good data fitting with the exception of fitting to X_3 data as expected (see Figure 4.6). Like data noise, missing data led to increased inaccuracy of the parameter estimates, regardless of the estimation methods. Finally, the computational speedup by using the incremental over the simultaneous estimation was significant, but was lower than in the previous runs due to the additional integrations of $\mathbf{X}_U$ and the larger number of independent parameters. The detailed values of the parameter estimates in this case study can be found in the Appendix B (Tables B2 and B3).

Table 4.3. Parameter estimations of the branched pathway model using noise-free data with X_3 missing.

	Simultaneous method		**Incremental method**	
	$\min \Phi_C$ [a]	$\min \Phi_S$	$\min \Phi_C$	$\min \Phi_S$
Computational time (sec)	85.03 h	4002.01 ± 696.11	1404.22 ± 120.71	445.47 ± 35.94
Average parameter error	71.90%	43.50% ± 2.34%	68.85% ± 4.57%	40.47% ± 0.59%
$\sqrt{\Phi_C}$	$\underline{4.54\times10^{-3}}$	6.46×10^{-3} ± 4.08×10^{-4}	$\underline{3.38\times10^{-3}}$ ± 1.14×10^{-4}	5.94×10^{-3} ± 3.23×10^{-5}
$\sqrt{\Phi_S}$	1.03	$\underline{2.99\times10^{-2}}$ ± 3.82×10^{-4}	8.32×10^{-2} ± 4.04×10^{-3}	$\underline{2.94\times10^{-2}}$ ± 2.77×10^{-6}

a. Only one out of five runs completed with a relative improvement of the objective function below 1% between iterations. The rest did not converge within the 5-day time limit after iterating for 471, 435, 863 and 786 times. The corresponding Φ_C at termination were 4.99×10^{-2}, 4.92×10^{-2}, 1.17×10^{-2} and 1.57×10^{-2}, respectively.

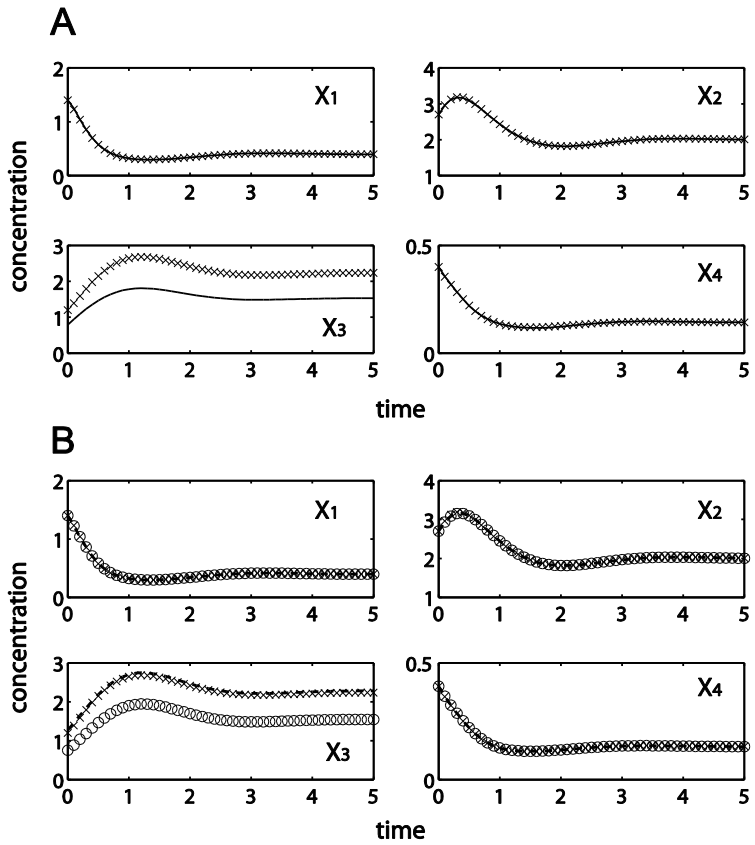


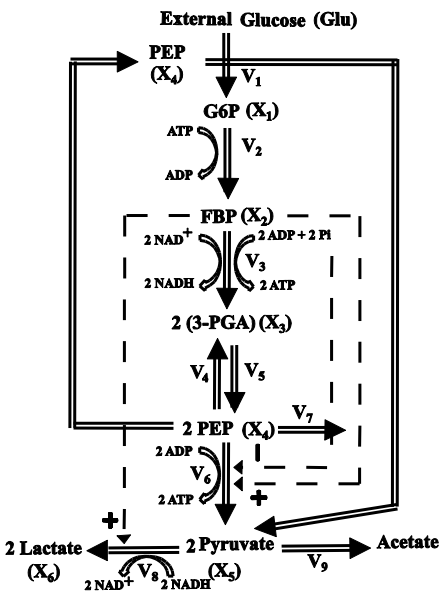
Figure 4.6. Simultaneous and incremental estimation of the branched pathway with missing X_3: *in silico* noise-free data (×). (A) Concentration predictions using parameter estimates from incremental method by Φ_C minimization (—); (B) Concentration predictions using parameter estimates from simultaneous method (○) and proposed method (- - -) by Φ_S minimization.

4.3.2 Glycolytic Pathway in *Lactococcus lactis*

The second case study was taken from the GMA modeling of the glycolytic pathway in *L. lactis* [78], involving six internal metabolites: glucose 6-phosphate (G6P) – X_1, fructose 1, 6-biphosphate (FBP) – X_2, 3-phosphoglycerate (3-PGA) – X_3, phosphoenolpyruvate (PEP) - X_4, Pyruvate – X_5, Lactate – X_6, and nine metabolic fluxes. External glucose (Glu), ATP and Pi were treated as off-line variables, whose values were interpolated from measurement data. The pathway

connectivity is given in Figure 4.7 A, while the model equations are provided in Figure 4.7 B with a total number of 25 rate constants and kinetic orders.

A



B

$$\begin{bmatrix} \dot{X}_1 \\ \dot{X}_2 \\ \dot{X}_3 \\ \dot{X}_4 \\ \dot{X}_5 \\ \dot{X}_6 \end{bmatrix} = \begin{bmatrix} 1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 2 & 1 & -1 & 0 & 0 & 0 & 0 \\ -1 & 0 & 0 & -1 & 1 & -1 & -1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 1 & 0 & -1 & -1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \\ v_3 \\ v_4 \\ v_5 \\ v_6 \\ v_7 \\ v_8 \\ v_9 \end{bmatrix}$$

$$v_1 = \gamma_1 Glu^{f_{1,Glu}} X_1^{f_{11}} X_4^{f_{14}}$$
$$v_2 = \gamma_2 X_1^{f_{21}} ATP^{f_{2,ATP}}$$
$$v_3 = \gamma_3 X_2^{f_{32}} P_i^{f_{3,Pi}}$$
$$v_4 = \gamma_4 X_4^{f_{44}}$$
$$v_5 = \gamma_5 X_3^{f_{53}}$$
$$v_6 = \gamma_6 X_2^{f_{62}} X_4^{f_{64}} P_i^{f_{6,Pi}}$$
$$v_7 = \gamma_7 X_4^{f_{74}}$$
$$v_8 = \gamma_8 X_5^{f_{85}} X_2^{f_{82}}$$
$$v_9 = \gamma_9 X_5^{f_{95}}$$

$$X_1(t_0) = 5.73$$
$$X_2(t_0) = 24.54$$
$$X_3(t_0) = 7.50$$
$$X_4(t_0) = 2.60$$
$$X_5(t_0) = 12.72$$
$$X_6(t_0) = 0.43$$

Figure 4.7. *L. lactis* glycolytic pathway: (A) Metabolic pathway map (Double-lined arrows: flow of material; dashed arrows with plus and minus signs: activation or inhibition, respectively) and (B) The GMA model equations [78].

The time-course concentration dataset of metabolites were measured using *in vivo* NMR [52,168], and smoothened data used for the parameter estimations below were shown in Figure 4.8. The raw data have been filtered previously [78], and these smoothened data for all metabolites but X_6, were directly used for the concentration-slope calculation in this case study. In the case of X_6, a saturating Hill-type equation: $k_1 t^n / (k_2 + t^n)$ where t is time and the constants k_1, k_2, n are

smoothing parameters, was fitted to the filtered data to remove unrealistic fluctuations. A central difference approximation was again adopted to obtain the time-slope data.

Fluxes v_4, v_7 and v_9 were selected as the independent set, again to give the least number of $\mathbf{p}_I$ and to ensure that $\mathbf{S}_D$ is invertible. All rate constants were constrained to within [0, 50], while the independent and dependent kinetic orders were allowed within [0, 5] and [-5, 5], respectively. The difference between the bounds for the independent and dependent kinetic orders was done on purpose to simulate a scenario where the signs of the independent kinetic orders were known *a priori*.

Table 4.4 reports the outcome of the single-step and incremental parameter estimation runs using Φ_C and Φ_S. The values of the parameter estimates are given in the Appendix B (Table B4). Like in the previous case study, there was a significant reduction in the estimation runtime by using the proposed method over the simultaneous estimation, with comparable goodness of fit in concentration and slope. None of the five repeats of Φ_C simultaneous minimization converged within the 5-day time limit, even after relaxing the convergence criterion of the objective function to 1%. On the other hand, the incremental estimation using Φ_C was not only able to converge, but was also faster than the simultaneous estimation of Φ_S that did not require any ODE integrations. The incremental estimation using Φ_C was able to provide parameters with the best overall concentration fit (see Figure 4.8), despite having a large slope error. Finally, minimizing Φ_S does not guarantee that the resulting ODE is numerically solvable,

as was the case of simultaneous estimation, due to numerical stiffness. But the incremental parameter estimation from minimizing Φ_S can produce solvable ODEs with good concentration and slope fits.

Table 4.4. Parameter estimations of the *L. lactis* model.

	Simultaneous method		**Incremental method**	
	$\min \Phi_C$ [a]	$\min \Phi_S$	$\min \Phi_C$	$\min \Phi_S$
Computational time (sec)	>5 days	3476.89 ± 349.63	976.72 ± 31.01	20.82 ± 2.71
$\sqrt{\Phi_C}$	—	Stiff ODE	<u>2.20</u> ± 8.81×10^{-3}	6.18 ± 7.28×10^{-2}
$\sqrt{\Phi_S}$	—	<u>2.67</u> ± 1.93×10^{-4}	1.51×10^{3} ± 52.50	<u>5.79</u> ± 9.62×10^{-4}

a. None of five runs finished with a relative improvement of the objective function below 1% within the 5-day time limit, after iterating for 60, 147, 93, 79 and 31 times. The corresponding Φ_C at termination were 9.31, 7.57, 8.77, 9.39 and 12.9, respectively.

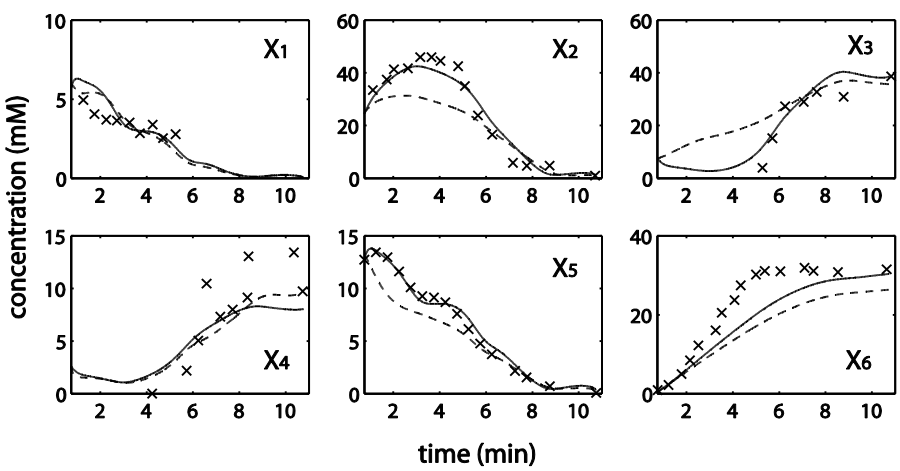


Figure 4.8. Incremental estimation of the *L. lactis* model: Experimental data (×) compared with model predictions using parameters from concentration error minimization (—) and slope error minimization (---).

4.4 Discussion

In this chapter, an incremental strategy is used to develop a computationally efficient method for the parameter estimation of ODE models. Unlike most commonly used methods, where the parameter estimation is performed to minimize model residuals over the entire parameter space simultaneously, here the estimation is done in two incremental steps, involving the estimation of dynamic reaction rates or fluxes and flux-based parameter regression. Importantly, the proposed strategy is designed to handle systems in which there exist extra degrees of freedom in the dynamic flux estimation, when the number of metabolic fluxes exceeds that of metabolites. The positive DOF means that there exist infinitely many solutions to the dynamic flux estimation, which is one of the factors underlying the parameter identifiability issues plaguing many estimation problems in Systems Biology [63,64].

The main premise of the new incremental method is in recognizing that while many equivalent solutions exist for the dynamic flux estimation, the subsequent flux-based regression will give parameter values with different goodness-of-fit, as measured by Φ_C or Φ_S. In other words, given any two dynamic flux vectors $\mathbf{v}(t_k)$ satisfying $\dot{\mathbf{X}}_m(t_k) = \mathbf{S}\mathbf{v}(t_k)$, the associated parameter pairs ($\mathbf{p}_I$, $\mathbf{p}_D$) may not predict the slope or concentration data equally well, due to differences in the quality of parameter regression for each $\mathbf{v}(t_k)$. In addition, because of the DOF, the minimization of model residuals needs to be done only over a subset of parameters that are associated with the flux degrees of freedom, resulting in much

reduced parameter search space and correspondingly much faster convergence to the (global) optimal solution. The superior performance of the proposed method over simultaneous estimation was convincingly demonstrated in the two GMA modeling case studies in the previous section. The minimization of slope error, also known as slope-estimation-decoupling strategy method [60], is arguably one of the most computationally efficient simultaneous methods. In this strategy, the parameter fitting essentially constitutes a zero-finding problem and the estimation can be done without having to integrate the ODEs. Yet, the incremental estimation could offer more than two orders of magnitude reduction in the computational time over this strategy.

As discussed in Chapter 2, there are many factors, including data-related, model-related, computational and mathematical issues, which contribute to the difficulty in estimating kinetic parameters of ODE models from time-course concentration data [24]. Each of these factors has been addressed to a certain degree by using the incremental identification strategy presented in this work. For example, in data-related issues, the proposed method can be modified to handle the absence of concentration data of some metabolites, as shown in Figure 4.2. Nevertheless, the method is neither able nor expected to resolve the lack of complete parameter identifiability due to insufficient (dynamical) information contained in the data [63,64]. As illustrated in the first case study, the single-step and incremental approaches did not provide any significant improvement in parameter estimate accuracy over the simultaneous method, and this accuracy expectedly deteriorated with noise contamination and loss of data.

The appropriateness of using a particular mathematical formulation, like power law, is an example of model-related issues. As discussed above, this issue can be addressed after the dynamic fluxes are estimated, where the chosen functional dependence of the fluxes on a specific set of metabolite concentrations can be tested prior to the parameter regression [77]. Next, the computational issues associated with performing a global optimization over a large number of variables and the need to integrate ODEs have been mitigated in the proposed method by performing optimization only over the independent parameter subset and using a minimization of slope error, respectively. Finally, in this work, we have also addressed a mathematical issue related to the degrees of freedom that exist during the inference of dynamic fluxes from slopes of concentration data. However, extra degrees of freedom (mathematical redundancies) are also expected to influence the second step of the method, i.e., one-flux-at-a-time parameter estimation. For (log)linear regression of parameters in the GMA models, such redundancy will lead to a lack of full column rank of the matrix containing the logarithms of concentration data $\mathbf{X}_m(t_k)$ and thus, can be straightforwardly detected.

The proposed estimation method has several weaknesses that are common among incremental estimation methods. As demonstrated in the first case study, the accuracy of the identified parameter relies on the ability to obtain good estimates of the concentration slopes. Direct slope estimation from the raw data, for example using central finite difference approximation, is usually not advisable due to high degree of noise in the typical biological data. Hence, pre-smoothing of

the time-course data is often required, as done in this study. Many algorithms are available for such purpose, from simplistic polynomial regression and splines to more advanced artificial neural network [59,60] and Whittaker-Eilers smoother [61,169]. If reliable concentration-slope estimates are not available, but bounds for the slope values can be obtained, then one can use interval arithmetic to derive upper and lower limits for the dependent fluxes and parameters using Equation 4.3 (or Equation 4.7) [170]. When the objective function involves integrating the model, validated solution to ODEs with interval parameters can be used to produce the corresponding upper and lower bounds of concentration predictions [171]. Finally, the estimation can be reformulated, for example by minimizing the upper bound of the objective.

In addition to the drawbacks discussed above, the proposed strategy requires *a priori* knowledge about the network topology, which requires complete information of the involved species, reaction stoichiometry and regulatory effects. Each aspect can pose significant challenges. For instance, an unknown species could be erroneously neglected in the pathway and may not be measured. Even with available measurements, the connectivity of this species with the others (by reaction or regulation) has to be strong enough to be retrieved from time-series concentration data, which can be done using methods including Bayesian network inference, transfer entropy, and Granger causality [172-174]. The inference of weak connectivity is challenging, as such issue relates to the fundamental problem of identifiability. I have discussed some of these problems in Section 3.4 Paragraph 5 and 6. Fortunately, one can usually gather some basic structure

information on the pathways of interest from existing publications and databases. For cellular metabolism, such information has become more readily available as genome-scale metabolic network of many important organisms, including human, *E. coli* and *S. cereviseae*, have been and are continuously being reconstructed [175].

CHAPTER 5 : ENSEMBLE KINETIC MODELING OF METABOLIC NETWORKS FROM DYNAMIC METABOLIC PROFILES

5.1 Summary

An incremental approach is used here for ensemble modeling of kinetic metabolic models. The model ensemble captures the uncertainty in the model identification procedure, specifically associated with the degree of freedom in the determination of metabolic fluxes from time profiles and non-identifiability of parameters from dynamic fluxes. The ensemble modeling method applies an existing parameter sampling strategy to explore and generate the viable parameter subspace, containing the model parameters that give statistically equivalent goodness of fit to metabolite time profiles. Built on the concept of incremental identification, the proposed ensemble modeling procedure relies on three components: (1) data smoothing and approximation of time-series metabolic concentration data, (2) a compact parameter space defining the model ensemble, and (3) efficient parameter exploration. The key contribution of this chapter lies in the use of an incremental approach to the building of model ensemble, making this process much more efficient and possible for applications to large-scale metabolic networks. The shift toward using a model ensemble, instead of the theoretically non-existent “best-fit” model, is necessary, as predictions from such best-fit model can be misleading. The performance of this ensemble modeling

approach has been demonstrated using the models of a generic branched pathway and the trehalose pathway in *Saccharomyces cerevisiae*.

5.2 Method

5.2.1 Problem Formulation

The incremental identification approach is again adopted to develop the model identification method in this chapter, but for different purpose and outcome. As described in Chapter 2, the estimation of dynamic fluxes from metabolite concentration slopes is commonly an underdetermined problem. In the previous chapter, the extra DOF in this estimation was used to restrict the parameter search space. However, the DOF also implies that there exist infinitely many dynamic flux combinations $\mathbf{v}(t_k)$ that can satisfy the mass balance equation $\dot{\mathbf{X}}_m(t_k) = \mathbf{S}\mathbf{v}(t_k)$, containing the true solutions. The set of such fluxes represent in essence the uncertainty that arises due to the lack of identifiability of dynamic fluxes from the time-course data. Because of such uncertainty, the true solution to the inverse modeling cannot be a single model, but rather an ensemble of models.

The focus of the new methodology development is therefore to construct this ensemble of models. Each of the models in this ensemble is derived from the same model equations and dynamic metabolic profiles, and member models differ only in the values of their kinetic parameters $\mathbf{p}$. In this case, the membership to the ensemble is tied to $\dot{\mathbf{X}}_m(t_k) = \mathbf{S}\mathbf{v}(t_k)$, where each model can provide (statistically) equivalent goodness of fit to the given dynamic metabolic profiles. The method follows the same procedure as the estimation strategy in Chapter 4, but naturally differs in the construction of the solution. First, given time-course

concentration data $\mathbf{X}_m(t_k)$, their time-slopes are estimated (e.g., using central difference approximation) after data smoothing. The discrepancy between the two methods begins from the calculation of the dynamic fluxes. Like before, the fluxes are decomposed into two groups: $\mathbf{v}(t_k)=[\ \mathbf{v}_I(t_k)^T\ \ \mathbf{v}_D(t_k)^T\ ]^T$, corresponding to the independent and dependent sets, respectively. The stoichiometric matrix **S** and the parameter vector **p** are also structured accordingly into $\mathbf{S} = [\ \mathbf{S}_I\ \ \mathbf{S}_D\ ]$ and $\mathbf{p} = [\ \mathbf{p}_I\ \ \mathbf{p}_D\ ]$. The dimension of the independent fluxes is given by the DOF, which is the difference between the number of fluxes and the number of metabolites: $n_{DOF} = n - m > 0$ (assuming the rank of **S** is equal to m). As stated in Chapter 4, given the values of the independent fluxes, the dependent fluxes can be computed according to Equation 4.3. This is the point where the two methods diverge.

The construction of the model ensemble is equivalent to the mapping of the viable region(s) in the parameter space, for which the corresponding model predictions fit the given data (statistically) equally well. Briefly, the method relies on an exploration of the independent parameter space to identify the viable regions for which (1) fluxes and parameters are within biologically relevant bounds and (2) the model prediction error is within acceptable statistical bounds. Here, the parameter space exploration is carried out using the HYPERSPACE toolbox, which implements an out-of-equilibrium adaptive Metropolis Monte Carlo (OEAMC) method and a multiple ellipsoid-based sampling (MEBS) method [176]. This toolbox was chosen as it has been shown to be very effective in exploring high-dimensional, non-convex and poorly connected viable spaces. Detailed steps of the ensemble construction are provided below.

5.2.2 HYPERSPACE Toolbox

The HYPERSPACE toolbox provided two key algorithms: the OEAMC and the MEBS methods. In short, the OEAMC method provides a coarse-grained global exploration of the viable parameter space. This coarse-grained set in turn becomes starting points for a fine-grained local exploration offered by MEBS to further characterize the space [176].

OEAMC Method

This algorithm was developed from Metropolis Monte Carlo sampling [177] and Simulated Annealing [150]. Given an initial viable parameter point, for example the parameter estimates $\mathbf{p}_I$ from the estimation method in Chapter 4, the OEAMC carries out n iterations in which new parameter points are sampled from a normal distribution and subjected to the criteria that define the desired viable region (see the next section). Specifically, the parameter space exploration starts from this known viable parameter point, around which the samples of the normal distribution are generated. A criterion manipulated by parameter β determines which point of the generated samples becomes the next sampling centre and influences the transition frequency between two parameter samples. This scheme is repeated for a predefined number of iterations n to guarantee that the defined whole space including disconnected areas could be sampled. After every n iterations, the algorithm determines whether the sampling should be continued depending on a convergence condition. Then, the viable parameter points (blue points in Figure 5.1) found so far are grouped into clusters and hyper-ellipsoids

(grey areas in Figure 5.1) with minimal volumes are constructed to enclose the viable points in each cluster. The convergence of this algorithm is determined from the sum of the volumes of these ellipsoids. The output is the set V_{MC} containing all the viable parameter points. Figure 5.1 illustrates the working of this algorithm.

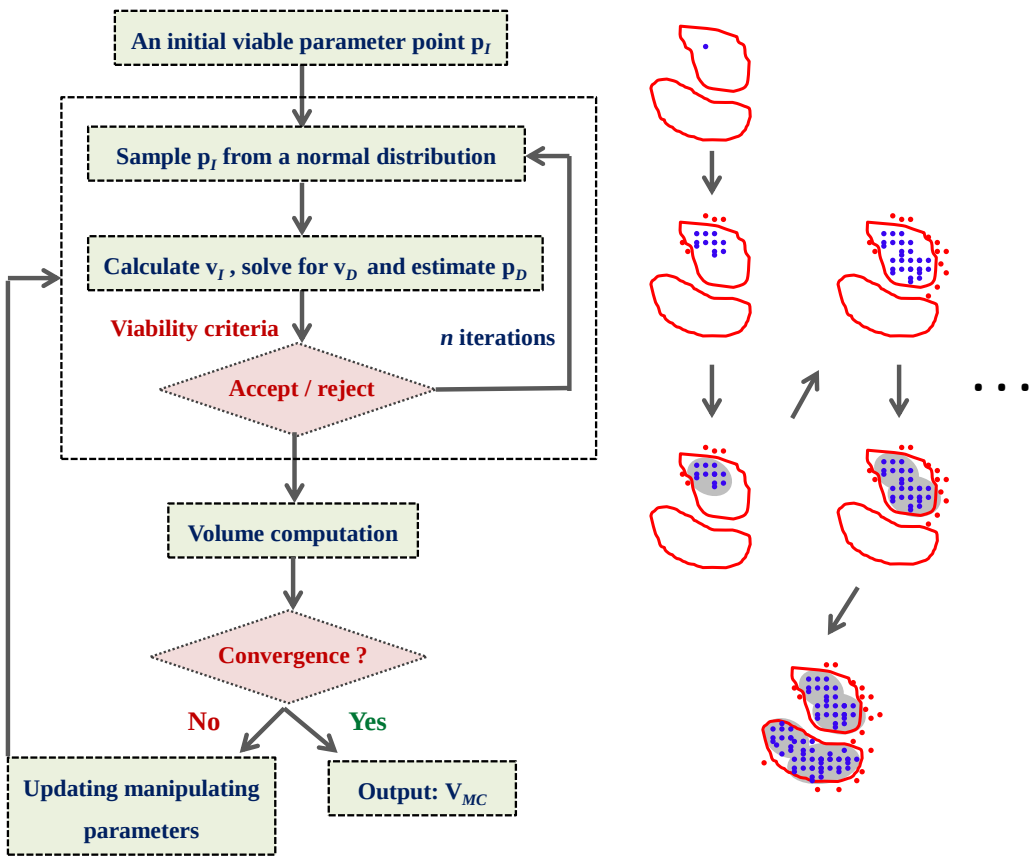


Figure 5.1. Flowchart of the OEAMC algorithm. On the right, the red closed curves represent hypothetical contour plots of the viable parameter space defined by some criteria. The viable points are marked in blue and the nonviable points are marked in red. Finally, the grey areas illustrate the minimum volume enclosing ellipsoids [176].

MEBS Method

The MEBS method produces hyper-ellipsoids to bound viable regions in the parameter search space, based on another algorithm that has been introduced elsewhere [178]. The ellipsoids' centers, orientations and lengths of axes can be fine tuned in order to enclose multiple viable regions as tightly as possible.

Starting from one parameter point of the set V_{MC} from the OEAMC algorithm, this method searches for viable parameter points near the boundary of the viable region. Then, it computes the Minimum Volume Enclosing Ellipsoid (MVEE, dashed curves in Figure 5.2) that covers these viable points, and samples inside an ellipsoid with the same orientation (solid curves in Figure 5.2) using larger axes scaled up by a multiplier g_i. Among the sample collection, the nonviable points (red points) are discarded, and based on the remaining viable ones (blue points), a new round of MVEE calculation and sampling are carried out with an updated g_{i+1}. The performance of the algorithm strongly depends on the multiplier g_i, and here I have used the recommended scaling parameters in the original publication [176]. The MEBS initiates a i+1-*th* ellipsoid expansion using the new sample point, which is chosen from the set composed by V_{MC} and the union (V_i) of viable points obtained after previous ellipsoid expansions. To explore the regions that have not been sampled, the algorithm preferentially selects a sample point that is far away from the average of all previous starting points. The iteration is repeated until the scaling multiplier trends to one or a fixed number of iterations is reached. Thereafter, another initial viable point is picked for another round of the above steps and this is repeated until the parameter points in the V_{MC} and V_i are

exhausted. The final output of the MEBS is a comprehensive set of viable parameter points. Figure 5.2 summarizes the procedure of the MEBS algorithm.

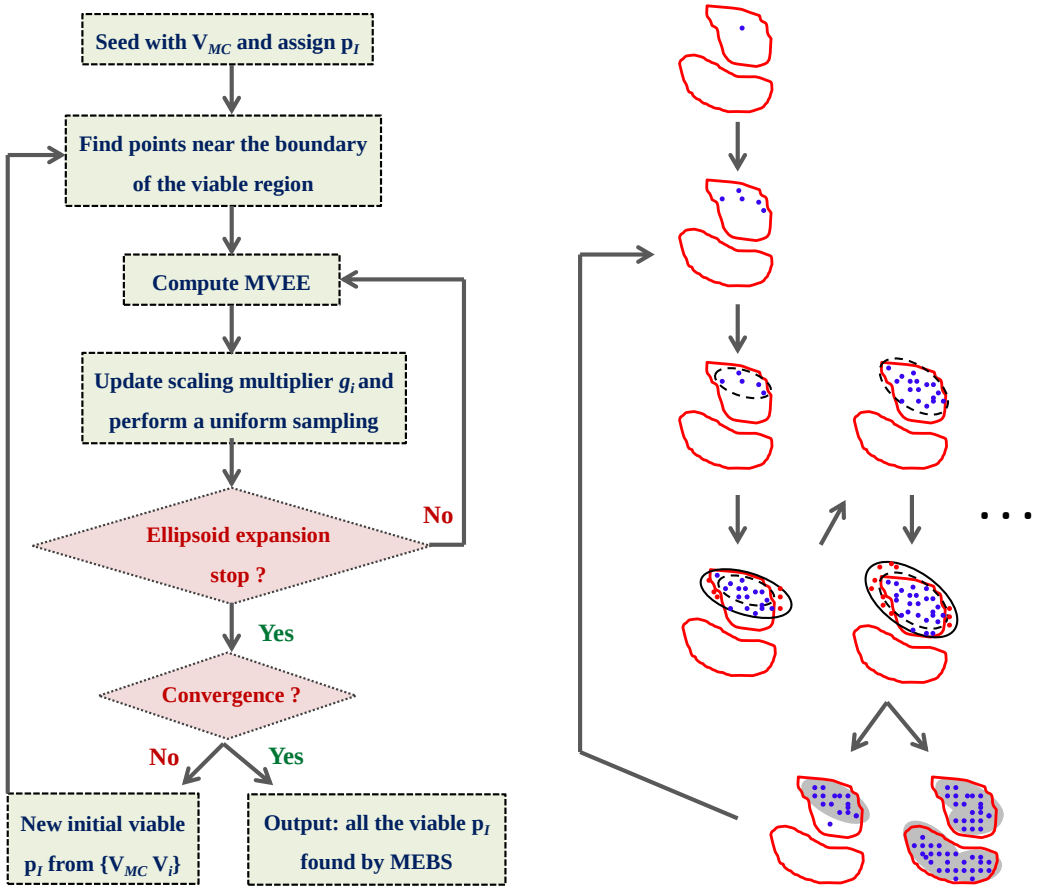


Figure 5.2. Flowchart of the MEBS algorithm. On the right, the red closed curves represent hypothetical contour plots of the viable parameter space defined by some criteria. The viable points are marked in blue and the nonviable points are marked in red. Finally, the grey areas illustrate the minimum volume enclosing ellipsoids [176].

One can apply the OEAMC and MEBS methods for a variety of purposes. In this case, the two algorithms will be used to characterize the space of parameters that defines the model ensemble. In the following sections, the criteria for viable parameters and the details of the application of the OEAMC and MEBS for the ensemble modeling will be provided. The HYPERSPACE toolbox also has an

inbuilt function to estimate the volume of the viable parameter space using Monte Carlo integration.

5.2.3 Parameter Bounds, Flux Bounds and Error Function Threshold

The first set of criteria for membership into the ensemble is related to the parameter and flux values, given by the following constraints:

$$\mathbf{p}_I \in [L_I, U_I]; \quad \mathbf{p}_D \in [L_D, U_D]; \quad \mathbf{v}_I(t_k) \in [0, U_v]; \quad \mathbf{v}_D(t_k) \in [0, U_v]; \quad (5.1)$$

where L_I (L_D) and U_I (U_D) denote the lower and upper bounds for the independent (dependent) parameters, and U_v is the maximum value of metabolic fluxes based on prior knowledge on the interested metabolic pathway. Reasonable bounds for rate constants and kinetic orders as well as the maximal value for metabolic fluxes in some specific organisms have been discussed previously in Chapter 2.

The second viability criterion is meant to establish equivalence among the member models in terms of their goodness of fit to data. If one makes the assumption that data noise comes from a Gaussian distribution, then the confidence bound of error function can be calculated using standard statistical analyses. Here, models with error function values within the confidence bound(s) are treated to be statistically equivalent. When data noise is not Gaussian and/or non-standard error function is used, the confidence interval can still be estimated using a Monte Carlo approach [179]

In this chapter, the confidence bound for the error function was obtained using a Monte Carlo approach. Specifically, the parameter estimation as described in Chapter 4 was repeatedly applied to 100 randomly generated time profiles from a Gaussian distribution with the noisy/experimental time profiles as the mean values. The variance of the data noise was estimated from the residuals of the data smoothing procedure. For each dataset, the same data smoothing and slope calculation were performed and the corresponding parameter estimates were obtained by minimizing the error function (see below). The confidence bound was directly estimated from the set of 100 error function values. For example, the 95% upper confidence bound of the error function is approximated by the *4-th* largest error function in this set.

5.2.4 Ensemble Modeling Procedure

In the case studies below, the error function was set to be:

$$\Phi_R(\mathbf{p},\mathbf{X}) = \frac{1}{mK}\sum_{k=1}^{K}\left[\mathbf{v}_D(t_k) - \mathbf{v}_D\left(\mathbf{X}_m(t_k),\mathbf{p}_D\right)\right]^T\left[\mathbf{v}_D(t_k) - \mathbf{v}_D\left(\mathbf{X}_m(t_k),\mathbf{p}_D\right)\right], \quad (5.2)$$

where K is the total number of measurement time points. This error function is implemented in the last step of the incremental estimation, where the kinetic parameters are regressed from the dynamic flux estimates. Note that the optimization of this error function was actually done one flux at a time. Of course, other error functions can be used, such as those given in Equations 4.4 and 4.5.

The model ensemble procedure starts with finding an initial viable point for the OEAMC algorithm, as discussed above. Next, the upper bound for the error function will be set either by applying standard statistical analysis assuming Gaussian noise or using the Monte Carlo algorithm described in the previous subsection. The OEAMC is then applied to generate the coarse-grained set of viable parameters over the space of the independent parameters. Finally, this set becomes the input to the MEBS algorithm, producing a population of viable parameters $\mathbf{p}_I$ that represents the ensemble of models. Note that while this work concerns with power-law fluxes, the ensemble generation procedure has general applicability to any kinetic models that can be written as $\dot{\mathbf{X}}(t,\mathbf{p}) = \mathbf{S}\mathbf{v}(\mathbf{X},\mathbf{p})$. The overall flowchart of the proposed ensemble modeling method can be summarized in Figure 5.3.

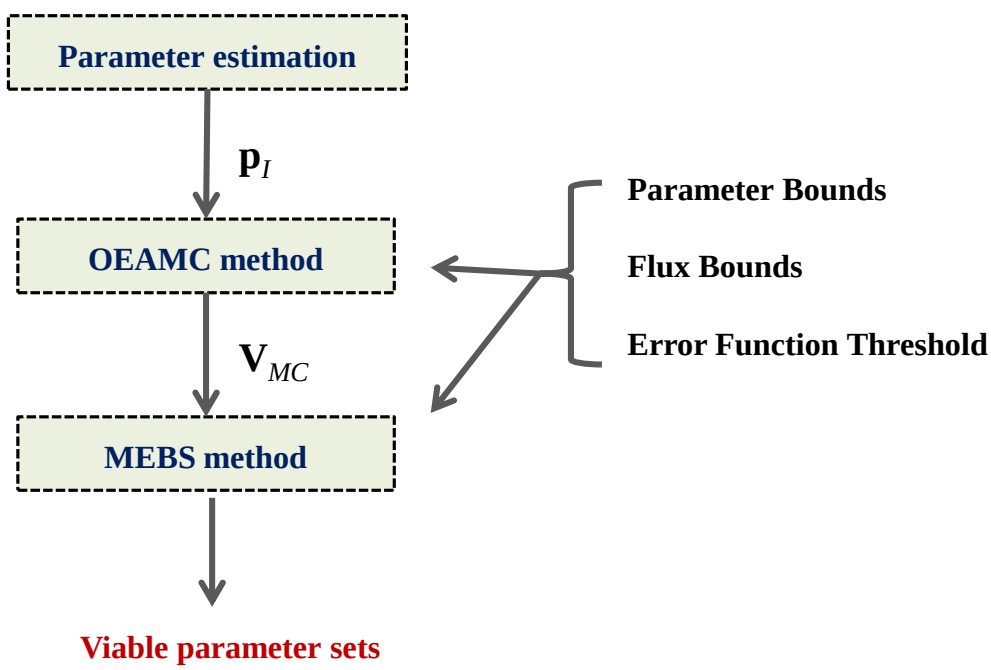


Figure 5.3. Flowchart of the proposed ensemble modeling method.

5.3 Results

The performance of the proposed procedure is demonstrated in the applications to two examples of GMA kinetic modeling, involving: a generic branched pathway [60] and the trehalose pathway in *Saccharomyces cerevisiae* [80].

5.3.1 A Generic Branched Pathway

The generic branched pathway in this example is the same as the one used for the first case study in Chapters 3 and 4. The pathway map and the GMA model equations are given in Figure 4.3. Like before, using the reported parameter values and initial concentrations [60] this model was used to generate *in silico* noisy time-course concentration data (i.i.d. Gaussian noise with zero mean and 10% standard deviation). For validation purpose, two independent datasets were generated in the same manner as above, but with different initial conditions $\left[X_1(t_0) \quad X_2(t_0) \quad X_3(t_0) \quad X_4(t_0)\right] = \left[4 \quad 1 \quad 3 \quad 4\right]$ and $\left[0.2 \quad 0.3 \quad 4.2 \quad 0.01\right]$, respectively. The noisy data were also smoothened using a *6*-th order polynomial, which provided the relatively best fit among polynomials according to adjusted R^2 [161] and Akaike Information Criterion (AIC) [167]. Subsequently, a central finite difference approximation was applied to compute the time-slopes of the smoothened noisy data.

The smoothened data were used to compute the initial parameter point, by way of applying the estimation method in Chapter 4 to minimize the regression

error $\sqrt{\Phi_R}$. The fluxes v_1 and v_6 were chosen as the independent set since this selection led to an invertible $\mathbf{S}_D$ and the least number of independent parameters. The independent parameters $\mathbf{p}_I$ included the rate constants $\{\gamma_1, \gamma_6\}$ and the kinetic orders $\{f_{13}, f_{64}\}$. The rate constants were constrained to within [0, 100], and the kinetic orders to within [0, 5] and the upper bound for allowable metabolic fluxes in this artificial network was set as 5×10^5 mM/min. The result of this estimation is summarized in Table 5.1. A comparison to Table 4.2 indicated that the minimization of $\sqrt{\Phi_R}$ can provide similar slope and concentration fittings to the objective functions used in Chapter 4. Finally, using the procedure described in Section 5.2.3, the upper confidence bound for the error function $\sqrt{\Phi_R}$ was determined to be 3.473×10^{-1}.

Table 5.1. Parameter estimation of the branched pathway model using Φ_R.

$\sqrt{\Phi_R}$	1.298×10^{-1}
$\sqrt{\Phi_C}$ [a]	4.125×10^{-2}
$\sqrt{\Phi_S}$ [b]	1.444×10^{-1}

a. Concentration error was calculated by Equation 4.4.

b. Slope error was calculated by Equation 4.5.

Table 5.2 summarizes the outcome of the ensemble modeling using a sequential application of the OEAMC and MEBS methods of the HYPERSPACE toolbox. The actual viable subspace of the independent parameters represented

only 0.284 % of the original space defined by the upper and lower parameter bounds. The ranges of concentration and slope errors were determined by generating a uniformly random sampling of parameters from the viable space (n = 59928) using Equations 4.4 and 4.5, respectively. Figure 5.4 shows two-dimensional projections of the viable regions onto the parameter axes of each independent flux. The member models of the ensemble were able to simulate the concentration and slope profiles reasonably well (see Table 5.2), as illustrated by the comparison of data and model predictions from five randomly selected models in the ensemble in Figure 5.5.

Table 5.2. Ensemble kinetic modeling of the branched pathway model using Φ_R.

Computational time (sec) [a]	1664
Calculated volume of initial parameter space (V_{ci}) [b]	2.5×10^{5}
Estimated volume of viable parameter space (V_{ev}) [c]	710.1 ± 5.1
Ratio of V_{ev} to V_{ci}	$(284.0 \pm 2.0) \times10^{-3}$ %
Value range of concentration errors $\sqrt{\Phi_C}$ [d]	$[3.554\times10^{-2}, 2.150\times10^{-1}]$
Value range of slope errors $\sqrt{\Phi_S}$ [e]	$[1.370\times10^{-1}, 5.081\times10^{-1}]$

a. The computational time was the total time of ensemble construction including OEAMC and MEBS phases, based on Dual Processors Intel Quad-Core 2.83 GHz.

b. V_{ci} was calculated through multiplication of initial parameter search ranges (i.e., 100×5×100×5).

c. V_{ev} was calculated by integrating the volumes of an ensemble of ellipsoids that cover the viable parameter space [176].

d. Concentration errors were calculated by Equation 4.4, given the parameter samples within the viable parameter space.

e. Slope errors were calculated by Equation 4.5, given the parameter samples within the viable parameter space.

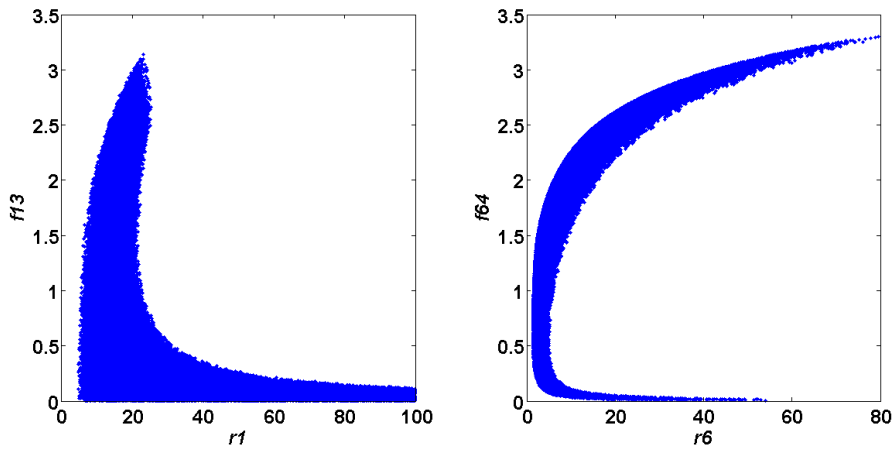


Figure 5.4. Two-dimensional projections of the viable parameter space onto the parameter axes of each independent flux (v_1: left, v_6: right).

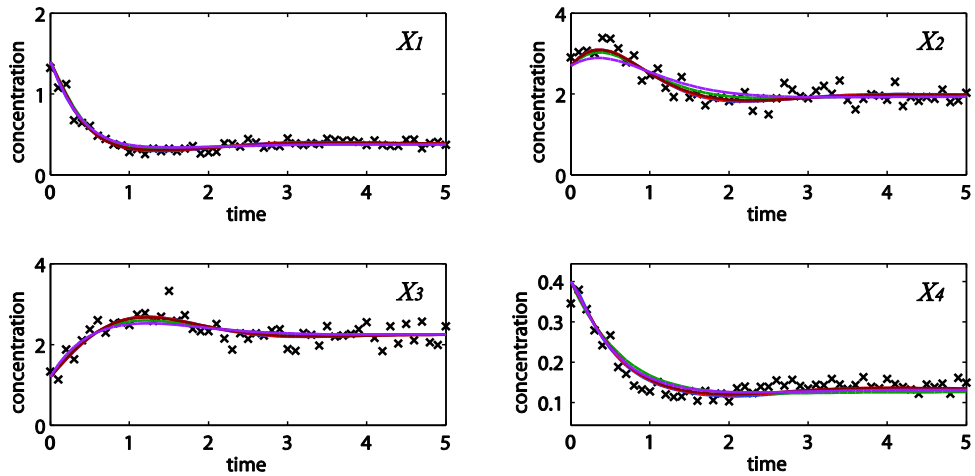


Figure 5.5. Concentration simulations of five randomly selected models from the ensemble (solid blue, brown, green, red and purple lines) versus the noisy data (×).

Finally, for model validation, Figure 5.6 shows the comparison of model simulations from the same five models and independent (simulated) experimental

datasets, indicating that these models could predict the systems dynamics under different initial conditions reasonably well.

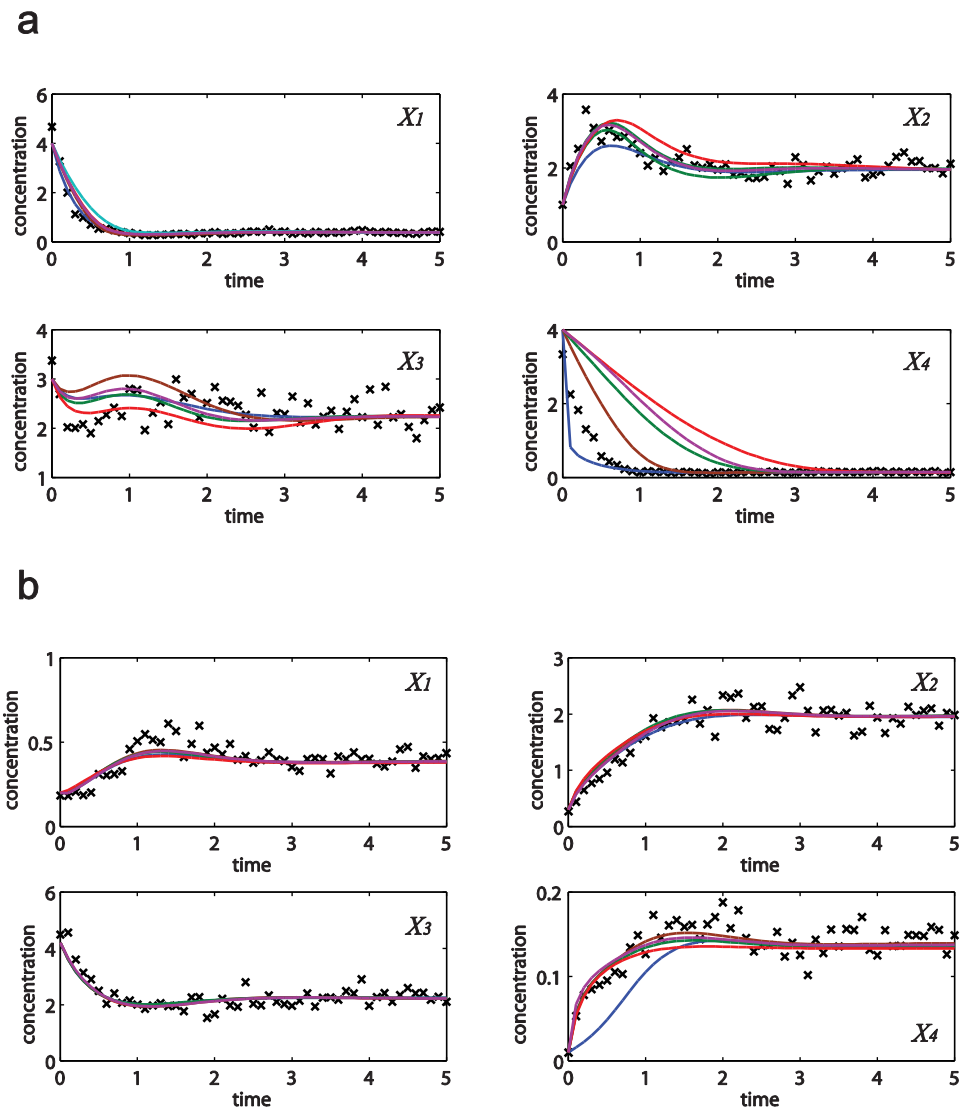


Figure 5.6. Concentration simulations of the same five models as in Figure 5.5 (solid blue, brown, green, red and purple lines) versus independent datasets (×), with initial concentrations of [4 1 3 4] (a) and [0.2 0.3 4.2 0.01] (b).

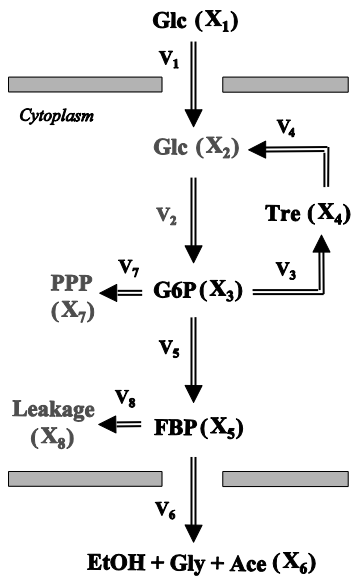
Note that besides the $\sqrt{\Phi_R}$ minimization, the proposed kinetic ensemble modeling approach can also incorporate other error functions, such as the slope

error $\sqrt{\Phi_S}$. The viable parameter space using the slope error $\sqrt{\Phi_S}$ closely resembles that shown in Figure 5.4 (see Appendix C), indicating that the robustness of the procedure in capturing the parametric uncertainty.

5.3.2 Trehalose pathway in *Saccharomyces cerevisiae*

The second case study was taken from the modeling of the glycolysis and trehalose production in the baker's yeast *Saccharomyces cerevisiae*. Figure 5.7 shows the metabolic pathway and the GMA model, which describes in a simplified fashion how glucose is converted into end products and how trehalose is synthesized and degraded in a cyclic pathway [79]. The concentrations of metabolites in this pathway are denoted as follows: extracellular glucose (Glc) – X_1, intracellular glucose (Glc) – X_2, glucose 6-phosphate (G6P) – X_3, trehalose (Tre) - X_4, fructose 1, 6-biphosphate (FBP) – X_5, extracellular end-products (ethanol, glycerol and acetate) – X_6, pentose phosphate pathway (PPP) – X_7 and other pathways (Leakage) – X_8. In this case, the time-course concentration data using *in vivo* NMR were only measured for X_1, X_3, X_4, X_5 and X_6 [80]. As an illustration here, we adopted the experimental dataset from normally grown cells at 30 °C that were fed with a pulse of glucose (see Figure 5.9). The raw experimental data were smoothened using a piecewise cubic spline, the fitting of which was validated by adjusted R^2 [161] and Akaike Information Criterion (AIC) [167]. Like before, a central difference approximation was applied to obtain the time-slopes of concentration data.

A



B

$$\begin{cases} \dot{X}_1 = -v_1/V_{ex} \\ \dot{X}_2 = (v_1 + 2v_4 - v_2)/V_{in} \\ \dot{X}_3 = (v_2 - 2v_3 - v_5 - v_7)/V_{in} \\ \dot{X}_4 = (v_3 - v_4)/V_{in} \\ \dot{X}_5 = (v_5 - v_6 - v_8)/V_{in} \\ \dot{X}_6 = 2v_6/V_{ex} \end{cases}$$

$$v_1 = f_1(X_1); \quad v_2 = f_2(X_2);$$
$$v_3 = f_3(X_3); \quad v_4 = f_4(X_4);$$
$$v_5 = f_5(X_3); \quad v_6 = f_6(X_5);$$
$$v_7 = f_7(X_3); \quad v_8 = f_8(X_5).$$

Figure 5.7. The trehalose pathway in *Saccharomyces cerevisiae*. (A) Metabolic pathway map. (B) The ODE model equations [79].

The ODE model contains six species and eight fluxes, as shown in Figure 5.7 B. In this case, X_7 and X_8 are not tracked, as none of the metabolites of interest depends on their concentrations. The variables V_{ex} and V_{in} denote the extracellular (5.00×10^{-2} L) and intracellular (7.17×10^{-3} L) volumes of the bioreactor and the cell population, respectively. While the intracellular glucose X_2 was not measured, the information for its rate of change can be obtained from the other measured metabolites, by performing an overall mass balancing, as follows:

$$\dot{X}_2 = (-\dot{X}_1 \cdot V_{ex} - \dot{X}_3 \cdot V_{in} - 2\dot{X}_4 \cdot V_{in} - \dot{X}_5 \cdot V_{in} - \tfrac{1}{2}\dot{X}_6 \cdot V_{ex} - v_7 - v_8)/V_{in} \qquad (5.3)$$

Using this relationship, the model was reduced to the following equations:

$$\begin{cases} \dot{X}_1 = -v_1/V_{ex} \\ \dot{X}_3 = (v_1 + 2v_4 - 2v_3 - v_5 - v_7)/V_{in} - \dot{X}_2 \\ \dot{X}_4 = (v_3 - v_4)/V_{in} \\ \dot{X}_5 = (v_5 - v_6 - v_8)/V_{in} \\ \dot{X}_6 = 2v_6/V_{ex} \end{cases} \qquad \begin{aligned} v_1 &= V_{\max 1} X_1 / \left(K_{\mathrm{m}1} + X_1\right) \\ v_3 &= \gamma_3 X_3^{f_{33}} \\ v_4 &= \gamma_4 X_4^{f_{44}} \\ v_5 &= \gamma_5 X_3^{f_{53}} \\ v_6 &= \gamma_6 X_5^{f_{65}} \\ v_7 &= \gamma_7 X_3^{f_{73}} \\ v_8 &= \gamma_8 X_5^{f_{85}} \end{aligned} \tag{5.4}$$

Hence, the estimation of the dependent fluxes was carried out without the need to integrate the ODE for X_2, which is different from the procedure outlined in Figure 4.2. The parameter estimation procedure in Figure 4.1 was applied to this reduced model.

Fluxes v_4, v_7 and v_8 were chosen as the independent fluxes according to the same criteria as before. In this case, v_7 and v_8 were associated with the unmeasured metabolite $\dot{X}_2$ according to Equation 5.3. Consequently, the independent parameters $\mathbf{p}_I$ comprise the rate constants $\{\gamma_4, \gamma_7, \gamma_8\}$ and the kinetic orders $\{f_{44}, f_{73}, f_{85}\}$, which were constrained within [0, 100] and [0, 5], respectively. Note that the glucose transport flux (v_1) was modeled using Michaelis-Menten (MM) kinetics instead of power law, as suggested from the time profile of X_1 (a constant decrease at high X_1 and an exponential-like time profile at low X_1). The regression of the MM parameters can also be casted as a linear regression problem as follows:

$$\begin{bmatrix} V_{\max 1} \\ K_{\mathrm{m}1} \end{bmatrix} = \left(\begin{bmatrix} X_1 & -v_1 \end{bmatrix}^T \begin{bmatrix} X_1 & -v_1 \end{bmatrix} \right)^{-1} \begin{bmatrix} X_1 & -v_1 \end{bmatrix}^T \begin{bmatrix} X_1 \cdot v_1 \end{bmatrix} \tag{5.5}$$

where $[X_1 \cdot v_1]$ is the vector product of element-wise multiplication of X_1 and v_1. Finally, the upper bound for flux values was set as 5×10^5 mM/min, according to the maximal flux value reported in a similar glycolytic pathway [125].

The initial parameter point to the OEAMC algorithm was again obtained by applying the parameter estimation procedure in Figure 4.1. Table 5.3 reports the result from this estimation. The same parameter estimation was applied repeatedly to 100 randomly generated datasets, again assuming Gaussian distributed noise with the experimental data as the mean values and variance that was estimated from the residuals of the smoothing procedure. In this case, the upper confidence bound for $\sqrt{\Phi_R}$ was estimated to be 1.860×10^{-1}.

Table 5.3. Parameter estimation of the trehalose pathway model using Φ_R.

$\sqrt{\Phi_R}$	7.639×10^{-2}
$\sqrt{\Phi_C}$	2.189
$\sqrt{\Phi_S}$	8.009

Table 5.4 gives the summary of the model ensemble for the trehalose model above. The volume of the viable region represents 2.590×10^{-3}% of the original constrained parameter space. The ranges of fitting error values were computed based on a uniform random sample of the viable parameter space (n = 3591) and Equations 4.4 and 4.5. Note that while the upper bound for the concentration errors was quite high, only a very small minority of the random parameter points

(3 out of 3591) had concentration errors than 10^2 and removing these, the upper bound for the concentration error reduces to 35.92. This issue is not completely unexpected as the model ensemble was created based on the flux error function and not the concentration error. In particular, there is no guarantee that parameter values with a small flux error will also provide a low concentration error. However, we note that the divergence between the flux error and concentration error functions occurred only rarely (< 0.1%). Figure 5.8 shows two-dimensional projections of the viable parameter subspace onto the parameter axes of each independent flux. A comparison between the concentration predictions by five randomly sampled models from the ensemble and the metabolite time profiles is shown in Figure 5.9.

Table 5.4. Ensemble kinetic modeling of the trehalose pathway model using Φ_R.

Computational time (sec)	6489
Calculated volume of initial parameter space (V_{ci}) [a]	1.25×10^8
Estimated volume of viable parameter space (V_{ev})	3237 ± 125
Ratio of V_{ev} to V_{ci}	$(25.90 \pm 1.00)\times10^{-4}$ %
Value range of concentration errors $\sqrt{\Phi_C}$	$[1.125, 3.880\times10^2]$
Value range of slope errors $\sqrt{\Phi_S}$	[5.825, 46.42]

a. V_{ci} was calculated through multiplication of initial parameter search ranges (i.e., 100×5×100×5×100×5).

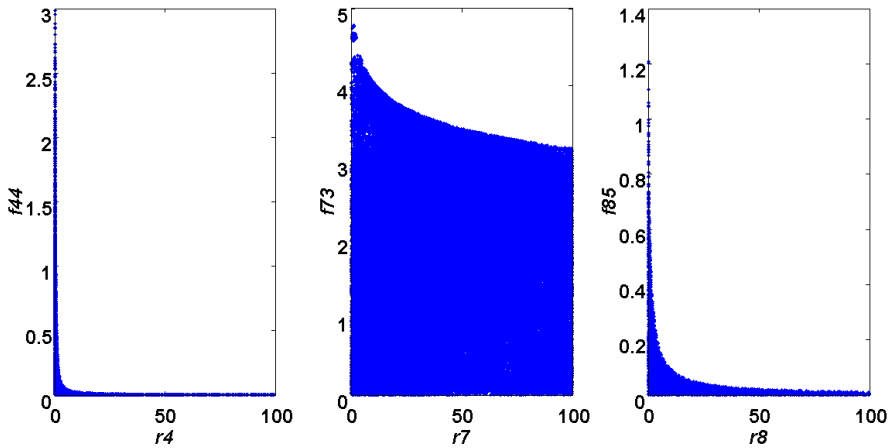


Figure 5.8. Two-dimensional projections of the viable parameter space onto the parameter axes of each independent flux (v_4: left, v_7: middle, v_8: right).

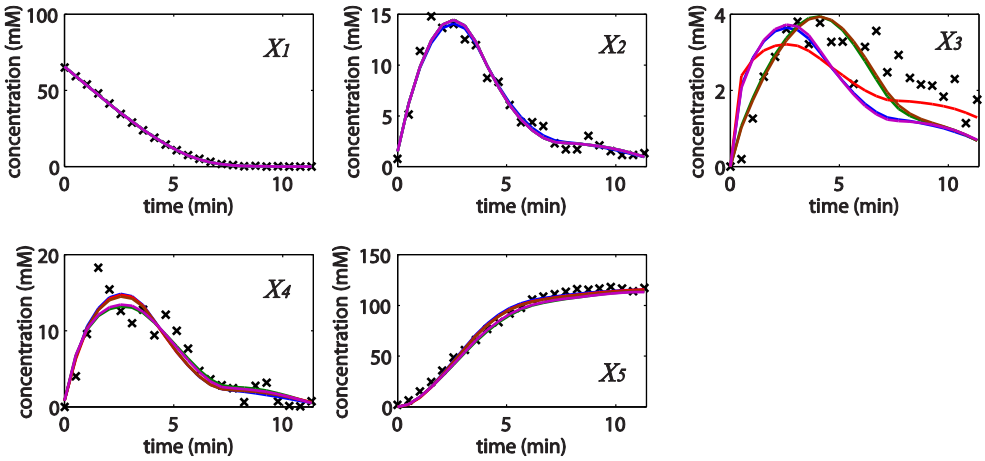


Figure 5.9. Concentration simulations of five randomly selected models from the ensemble (solid blue, brown, green, red and purple lines) versus the experimental data (×).

5.4 Discussion

The difficulty in simultaneously estimating kinetic parameters of metabolic models is often caused by the lack of complete parameter identifiability [64]. In other words, not all parameters can be uniquely identified and consequently many parameter combinations can give similar data fittings [160]. Hence, even if an estimation algorithm can return the "best-fit" model for a given dataset, this model may have little predictive capability, or worse, can be misleading. The model identification procedure in this chapter circumvents this problem by using an incremental identification approach to generate an ensemble of equivalent models in the sense that (1) the models closely approximate the same mass balance equation and (2) the model approximations are statistically equal (to within a 95% confidence level). Although the case studies mainly involved GMA models with power-law flux functions, the ensemble modeling procedure can be used for any form of flux functions, as long as the ODE model follows equation $\dot{\mathbf{X}}_m(t_k) = \mathbf{S}\mathbf{v}(t_k)$. For power-law and Michaelis-Menten kinetics, the least square regression of the dependent parameters reduces to linear regression, and thus can be done very efficiently. The main reason to use power-law models here was that they represent some of the most challenging problems in kinetic modeling due to the large parameter space, the lack of complete parameter identifiability, stiff ODEs and high degree of nonlinearity.

The proposed ensemble modeling method has the advantages that (1) the model ensemble is compactly defined using a small number of independent

parameters; (2) the dependent parameters can be efficiently computed from the independent parameters; (3) only biologically-meaningful models are included in the model ensemble; and (4) data uncertainty (noise) is explicitly accounted for. The first two aspects come as courtesy of the step-wise identification approach adopted in the development of the method. The computational cost of constructing the model ensemble is related with the parameter exploration and the computation of the error function. The compactness of the parameter space of the ensemble is therefore particularly important for numerical efficiency and ultimately for practical applications. For OEAMC and MEBS algorithms, the number of required parameter samples during parameter exploration has been shown to increase linearly with the parameter dimension, which in this case is equal to the number of independent parameters [176]. On the other hand, the computational cost of a single evaluation of the error function primarily comes from the least square regression of the dependent parameters and possibly from the integration of the ODE, if the error function requires the simulation of X(t). For the error function used in the case studies above, this computational cost should increase linearly with the number of dependent fluxes, assuming that the number of unknown parameters in each dependent flux stays about the same.

In the proposed ensemble modeling, the model uncertainty is related to parametric uncertainty that arises from data noise, leaving out the contribution of structural uncertainty (mismatch between the assumed model equations and the true dynamics). Increasing data noise is therefore expected to increase the size of the model ensemble, i.e. the volume of the viable parameter subspace, by directly

changing the statistics of the error function. However, in this case, higher noise in data will also lead to more uncertainty in the time slopes estimates of the concentration data. Since the direct (error function) and indirect (smoothing and slope calculation) effects of data noise could not be easily separated, we have chosen a Monte Carlo approach in determining the confidence bound of the error function (see Method section).

In the ensemble modeling, I have assumed that time-series data for all metabolites in the model are available. When one or more metabolites are not measured, before performing the proposed procedure, one can modify the procedure by first rewriting the ODE model as Equation (4.6), separating same as the balances associated with measured and unmeasured, and simulate the data of unmeasured ones in the same way as described in Section 4.2. In addition, I have made another assumption that there exists a unique solution to the computation of $\mathbf{p}_D$ from $\mathbf{p}_I$. For GMA models, this assumption requires that (1) the number of time points exceed the number of parameters $\mathbf{p}_D$ from each flux (not the total number) and (2) the logarithm of the metabolite concentration time profiles appearing in each flux are linearly independent and non-constant. The first requirement is usually satisfied as the number of parameters involved in every flux ranges only between 2 and 5. The second requirement depends on the experimental conditions, but is again usually fulfilled since each flux depends only on a handful of metabolites and data are contaminated with random noise. If this assumption becomes invalid for one or more dependent fluxes, then these fluxes can be included into the set of independent fluxes, at the cost of increasing

the dimensionality and computational time of the parameter exploration step. In such a case, the calculation of dependent fluxes from the independent flux values will require taking a pseudo-inverse of $\mathbf{S}_D$ (see Method).

Constraints on parameters and fluxes are important in restricting the size of the ensemble, in a problem dependent manner. For example, in the first case study, the ensemble hit the lower constraints on both kinetic order parameters (set at 0) and the upper constraint for the rate constant γ_1 (see Figure 5.4). Meanwhile, parameter constraints affect the second case study more than the first, where the lower and upper constraints of all rate constants and the lower bounds of all kinetic orders limited the viable parameter subspace (see Figure 5.8). Furthermore, in both case studies, the requirement for positivity of the flux values (i.e. lower bounds of the fluxes) was an important constraint, as this was frequently violated during the parameter exploration (data not shown).

While the kinetic ensemble models are considered equivalent, each will give a slightly different goodness of fit to the data that were used to identify them. This difference arises from two factors: (1) the least square regression of the dependent parameter values and (2) the use of concentration slopes. The former implies that the dependent fluxes are only approximations in the model, and the latter means that the mass balance is only satisfied at discrete time points (since the ODEs are not integrated in this procedure). In fact, such difference between the member models is expected and reasonable considering that the “best-fit” model to a single dataset is not reliable due to the data uncertainty and the lack of full model identifiability (existing DOFs).

Model building in systems biology, especially for metabolic networks, is usually formulated as an iterative procedure. While typically such a procedure considers a single “optimal” model, there is no guarantee, however, that the iteration will converge to a single model due to the issue of model identifiability, as mentioned before. The ensemble model creation in this chapter can be integrated in such an iterative procedure and here, the efficacy of the generation and screening of member models is important. Such efficiency can be guaranteed by the inbuilt features of the proposed method, including incremental identification (from time-slope approximation and dynamic flux calculation to linearized flux-based parameter estimation) and parameter space reduction (the independent parameter set). At each iteration, the ensemble size can be reduced by removal of member models that are not consistent with the additional experimental data. In this sense, any progress in accurate quantification of dynamic fluxes turns to be very helpful and is ready to be directly applied in the proposed method.

In addition, the ensemble of kinetic models can be further pruned using existing strategies in the generation of an ensemble of metabolic models. For example, steady-state data from knock-out studies and thermodynamic principles can be used as criteria for further reducing the size of generated ensemble models [180]. As another example, the benefits of improving the quantification of dynamic fluxes will immediately materialize as such data can be directly used in the proposed method. This resulted ensemble again allows for re-examining the possible phenotypes of the network upon new information or perturbations, such

as enzyme over-expression. This topic is currently under investigation within our group.

Finally, the ability to generate an ensemble of kinetic models also necessitates the development of new methodologies on how to utilize such ensemble. The obvious challenge is how to analyze and/or optimize the system when it is represented by a set of models, not just one model, which may contain an infinite number of members. Here, we propose two strategies: the first involves the generation of a (random) sample of models from the ensemble and in such a case, the results from the analysis and optimization can be represented in the form of a histogram. The second strategy is to take the advantage that the ensemble model generation involves only linear (or log-linear) algebraic equations. In this case, interval or constraint propagation using interval arithmetic can be used to evaluate upper and lower bounds for the system behaviors, as done previously for GMA models [106].

CHAPTER 6 : CONCLUSIONS AND FUTURE WORK

6.1 Conclusions

Advancements in biological techniques have made time-series measurements of metabolite concentrations more readily available, providing the necessary (though likely still insufficient) data to build kinetic models of cellular metabolism. Deciphering the information contained in these data about the structure and dynamics of metabolic pathways, is challenging but important. Kinetic metabolic models will provide an invaluable quantitative tool for metabolic engineering efforts [181]. Specifically, kinetic models such as those presented in this thesis, can be used to predict system responses to perturbations in either regulatory or metabolic networks. When coupled with an optimization procedure, the models can lend a hand in guiding genetic manipulations in pathway optimization.

The process of building kinetic models, however, is often complicated by issues related with data, model, computation and mathematics. Specially, some of the most challenging problems include the lack of complete data and poor data quality (noise), the high computational cost in performing parameter estimation from data, and the lack of complete parameter identifiability. Resolving these problems constitutes the main objective of this thesis, through the development of new estimation methodologies.

In Chapter 3, an incremental and iterative parameter estimation method was developed to address two issues: (1) missing metabolite time profiles and (2) high computational cost associated with integrating kinetic ODE models during parameter estimation. A new method was built from a combination of two existing strategies: the decoupling method and the ODE decomposition method. The decoupling method was known to be fast, as it requires no integration of the ODE models. However, the requirement of having a completely measured system reduces the practical significance of this method. The new combined method loses some of the computational efficiency of the decoupling method, as ODE integrations are still performed, in order to remove the above requirement. Hence, the computational performance of the new method straddles between the faster decoupling and the slower ODE decomposition methods.

However, the deeper issue in parameter estimation related to identifiability, especially given incomplete data, is still not accounted for in the iterative estimation method above. The lack of complete identifiability is often the real reason why existing algorithms fail in obtaining accurate parameter estimates. This issue becomes the focus of Chapters 4 and 5, in which I have taken an incremental identification approach [85,102]. In this approach, the parameter estimation of kinetic ODE models is decomposed into smaller sequential sub-problems. By doing so, the issue of identifiability can be addressed directly at each step.

In Chapter 4, a new incremental estimation method was formulated to address the degrees of freedom that arise from the lower the number of species than the

reactions. By using an incremental approach, this DOF could be addressed directly during the estimation of reaction fluxes from the metabolite concentration slopes. In this case, only a subset of fluxes is independent, i.e., given the values of these fluxes, the values of the remaining fluxes can be constrained by the mass balance. The important step forward was to recognize that the minimization of error function to estimate parameters can be done over the set of independent fluxes, offering a significant reduction in the optimization search space. In the case studies, the new incremental method outperformed the traditional simultaneous estimation methods by providing smaller errors in parameter estimates, slope and concentration fittings and importantly, by significantly reducing the computational time. The estimation method was flexible enough to handle cases of incomplete time profile data with little modifications.

Specifically, the improvement offered by the new incremental estimation can be attributed to three unique features of this method: (1) parameter estimation is restricted only within the flux-defined subspace (independent set), reducing complexity and computational effort greatly; (2) the incremental approach enables easy diagnosis of estimation errors during each step or increment with the flexibility of validating the assumed flux functions; and (3) the interplay between concentration and slope fittings helps to reduce the potential error compensations within the fluxes and to enhance accuracy in both slope and concentration predictions. These advantages will bring the kinetic modeling of genome-scale metabolic networks closer to reality.

In Chapter 5, the incremental identification approach was used to produce a completely different outcome. Here, a new method was developed to construct an ensemble of kinetic models of metabolic networks, where models in the ensemble have equivalent goodness of fit to given time profiles. The lack of complete parameter identifiability implies that the estimation problem is underdetermined and thus, there is no unique solution to the inverse problem. This corollary was precisely the motivation to generate such an ensemble. In essence, the ensemble recapitulates the parametric uncertainty arising from the parameter identifiability issue. In fact, the use of the "best fit" model, resulting from the traditional model identification procedure, maybe misleading.

Briefly, the proposed method was built using an efficient random sampling of viable parameter subspace [176] to generate the set of biologically meaningful values for the independent fluxes, based on which the dependent fluxes and the associated parameters can be obtained. During different steps in the construction, feasibility checks were performed to make sure that the calculated fluxes and parameters were within reasonable bounds. The method however only addressed two sources of uncertainty: (1) data noise and (2) the DOF in the estimation of fluxes as discussed above. Hence, this part of the thesis represents only a starting point of further investigations to incorporate other sources of uncertainty and importantly, to build new tools for applying such an ensemble in Metabolic Engineering.

6.2 Future Work

Although I have addressed several commonly encountered problems in the process of kinetic model identification, many challenges still exist, requiring more in-depth work. Below, I have outlined important and relevant research topics to the present thesis.

6.2.1 Data Smoothing

Biological measurements usually contain significant level of noise, which complicates the application of the methods relying on estimating the time-slopes of such data. In this regard, reliable methods for data smoothing are highly desirable. During the investigations in this thesis, a few algorithms have been tried, such as splines [54-56], polynomial fitting [57], filters [58], artificial neural networks (ANNs) [59,60] and AutoSmoother [61], but the success of each algorithm varies on a case-by-case basis.

The “ideal” smoother however should provide reliable performance, giving unbiased estimates of the slopes, regardless of the noise structure within the data. This aspect of the methods needs to be addressed more carefully and deserves deeper investigations. A case study in Chapter 4 (branched pathway) indicated that up to 70% of the parameter error arose from the inaccuracy in the estimation of time-slopes of concentration data (comparing Tables B1 and 4.2), when measurements of all metabolites are available. One possible strategy is to use a

hybrid of algorithms, for example, by taking an average of the estimates of time-slopes from the algorithms above.

6.2.2 Ensemble Kinetic Modeling in Consideration of Model Uncertainty

Many difficulties within the inverse modeling from data are rooted from the fundamental issue of model uncertainty. Here, model uncertainty accounts for most of the uncertain factors in the model identification process, including (1) structural uncertainty, (2) parametric uncertainty and (3) dynamic uncertainty (see Figure 6.1). As a result, efforts on finding a single (best-fit) model may become pointless due to the aforementioned uncertainties and the best-fit model, if found, may have little predictive capability or worse, it could be misleading [182].

Here, a different strategy, tackling each aspect of model uncertainty, is clearly needed, so that eventually it become possible to create a more comprehensive ensemble of kinetic models in consideration of all the three aspects. Those uncertainties can be recapitulated in the member models of the ensemble, which differ from each other (1) in the material (or information) flows of the biological pathways (structural uncertainty), (2) in the values of dynamic fluxes and kinetic parameters (parametric uncertainty), and (3) in the mathematical approximations of biochemical reactions (dynamic uncertainty). The methodology described in Chapter 5 addressed the second factor, i.e., the parametric uncertainty.

Structural uncertainty results from the lack of complete knowledge on the topology or connectivity in the pathways. In this case, a set of biologically feasible pathways could be constructed and used to generate the ensemble model, which could be mathematically described as:

$$\{\mathbf{S}\} = \{\mathbf{S} : \mathbf{S} = \mathbf{NP}\} \tag{6.1}$$

where **N** denotes a superset of stoichiometric matrix, including all the possible connections (reactions) in the network, **P** is a permutation matrix to remove unlikely links between components (metabolites), and **S** represents the confirmed topology. Such structural uncertainty can be further constrained using the methods of model-free or model-based structural identification, such as target factor analysis (TFA) [88].

Lastly, there also exists uncertainty in the formulation of the flux functions. In this regards, mechanistically based or canonical models, such as Michaelis-Menten, Hill-type and power-law kinetics, have been commonly used to describe the flux dynamics. Nevertheless, it may not be appropriate to adopt the same basis function for modeling all the fluxes that show distinctive characteristics. Hence, it is also important to tailor the modeling frameworks to accommodate the possibility of combining different canonical flux functions. Finally, the model ensemble should account all of these uncertainty sources, as illustrated in Figure 6.1.

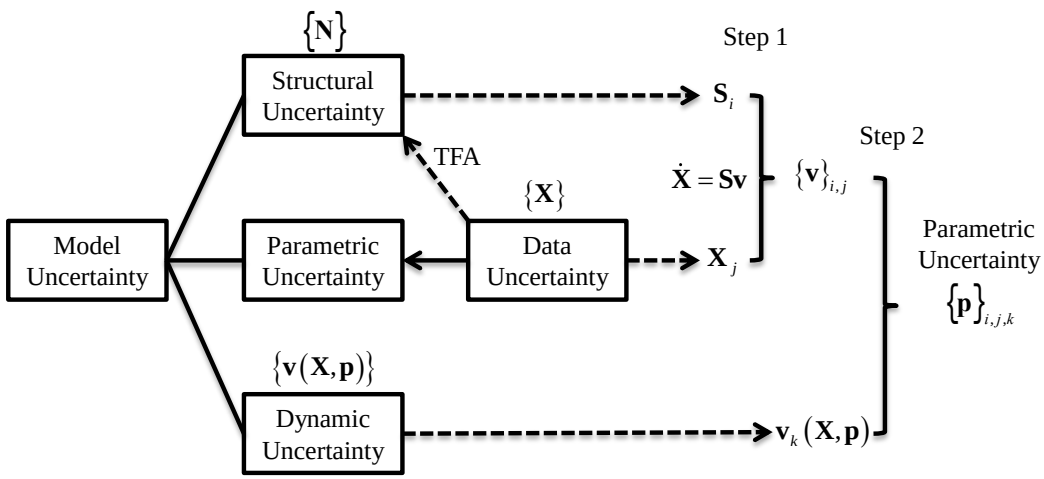


Figure 6.1. Model uncertainty and its parameterization, in which structural, dynamic and data uncertainties are represented by the sets {**N**}, {**v**(**X**,**p**)} and [153], respectively.

6.2.3 Applications of Ensemble of Kinetic Models

Along with the ensemble construction of kinetic models, the computational tools for the applications of the built models also need to be developed. The tools will need to consider (1) what type of outputs is desired from the model ensemble and (2) the computational requirement in producing these outputs from the ensemble. These two considerations will depend on the specific applications. For example, let us assume that the model ensemble is used to come up with the input profile (e.g., glucose) that optimizes some system performance (e.g., ethanol yield). In this case, one can borrow a concept from robust control theory, in which the ensemble model will be used to predict the worst-case performance among all feasible models, and the optimization of input profiles will be done to maximize this worst-case performance. One possible strategy to estimate the worst-case performance using the ensemble modeling is to use validated ODE solvers [171].

Finally, the future research topics discussed in this chapter: data smoothing, kinetic ensemble modeling and its applications, can and should still be integrated into the model identification cycle (Figure 1.3). The research findings shall provide the enabling tools for kinetic modeling under uncertainties and for resolving the issues related to data, model, computation and mathematics in the process of model identification of metabolic pathways as well as other biological networks.

BIBLIOGRAPHY

1. Otero, J.M.; Nielsen, J., Industrial systems biology. *Biotechnol Bioeng* **2010**, *105*, 439-460.

2. Stephanopoulos, G.; Aristidou, A.A.; Nielsen, J.H., *Metabolic engineering : Principles and methodologies*. Academic Press: San Diego, 1998; p xxi, 725 p.

3. Alper, H.; Miyaoku, K.; Stephanopoulos, G., Construction of lycopene-overproducing e. Coli strains by combining systematic and combinatorial gene knockout targets. *Nat Biotechnol* **2005**, *23*, 612-616.

4. Ro, D.K.; Paradise, E.M.; Ouellet, M.; Fisher, K.J.; Newman, K.L.; Ndungu, J.M.; Ho, K.A.; Eachus, R.A.; Ham, T.S.; Kirby, J., *et al.*, Production of the antimalarial drug precursor artemisinic acid in engineered yeast. *Nature* **2006**, *440*, 940-943.

5. Minami, H.; Kim, J.S.; Ikezawa, N.; Takemura, T.; Katayama, T.; Kumagai, H.; Sato, F., Microbial production of plant benzylisoquinoline alkaloids. *Proc Natl Acad Sci U S A* **2008**, *105*, 7393-7398.

6. Park, J.H.; Lee, K.H.; Kim, T.Y.; Lee, S.Y., Metabolic engineering of escherichia coli for the production of l-valine based on transcriptome analysis and in silico gene knockout simulation. *Proc Natl Acad Sci U S A* **2007**, *104*, 7797-7802.

7. Hunter, W.N., The non-mevalonate pathway of isoprenoid precursor biosynthesis. *J Biol Chem* **2007**, *282*, 21573-21577.

8. Stephanopoulos G. ; Aristidou A.A. ; J., N., *Metabolic engineering: Principles and methodologies*. Academic Press: San Diego, 1998; Vol. 1.

9. Bailey, J.E., Mathematical modeling and analysis in biochemical engineering: Past accomplishments and future opportunities. *Biotechnol Prog* **1998**, *14*, 8-20.

10. Palsson, B., *Systems biology : Properties of reconstructed networks.* Cambridge University Press: New York, 2006.

11. Llaneras, F.; Pico, J., Stoichiometric modelling of cell metabolism. *J Biosci Bioeng* **2008**, *105*, 1-11.

12. Gombert, A.K.; Nielsen, J., Mathematical modelling of metabolism. *Curr Opin Biotechnol* **2000**, *11*, 180-186.

13. Llaneras, F.; Pico, J., Stoichiometric modelling of cell metabolism. *J Biosci Bioeng* **2008**, *105*, 1-11.

14. Metallo, C.M.; Walther, J.L.; Stephanopoulos, G., Evaluation of (13)c isotopic tracers for metabolic flux analysis in mammalian cells. *J Biotechnol* **2009**, *144*, 167-174.

15. Varma, A.; Palsson, B.O., Metabolic flux balancing - basic concepts, scientific and practical use. *Bio-Technol* **1994**, *12*, 994-998.

16. Palsson, B., *Systems biology : Properties of reconstructed networks*. Cambridge University Press: Cambridge, 2006; p xii, 322 p.

17. Schuster, S.; Heinrich, R., Minimization of intermediate concentrations as a suggested optimality principle for biochemical networks .1. Theoretical-analysis. *J Math Biol* **1991**, *29*, 425-442.

18. Schuster, S.; Schuster, R.; Heinrich, R., Minimization of intermediate concentrations as a suggested optimality principle for biochemical networks .2. Time hierarchy, enzymatic rate laws, and erythrocyte metabolism. *J Math Biol* **1991**, *29*, 443-455.

19. Papin, J.A.; Stelling, J.; Price, N.D.; Klamt, S.; Schuster, S.; Palsson, B.O., Comparison of network-based pathway analysis methods. *Trends Biotechnol* **2004**, *22*, 400-405.

20. Schilling, C.H.; Palsson, B.O., The underlying pathway structure of biochemical reaction networks. *P Natl Acad Sci USA* **1998**, *95*, 4193-4198.

21. Schilling, C.H.; Letscher, D.; Palsson, B.O., Theory for the systemic definition of metabolic pathways and their use in interpreting metabolic function from a pathway-oriented perspective. *J Theor Biol* **2000**, *203*, 229-248.

22. Schuster, S.; Fell, D.A.; Dandekar, T., A general definition of metabolic pathways useful for systematic organization and analysis of complex metabolic networks. *Nat Biotechnol* **2000**, *18*, 326-332.

23. Schuster, S.; Hilgetag, C.; Woods, J.H.; Fell, D.A., Reaction routes in biochemical reaction systems: Algebraic properties, validated calculation procedure and example from nucleotide metabolism. *J Math Biol* **2002**, *45*, 153-181.

24. Chou, I.C.; Voit, E.O., Recent developments in parameter estimation and structure identification of biochemical and genomic systems. *Math Biosci* **2009**, *219*, 57-83.

25. Schulz, A.R., *Enzyme kinetics: From diastase to multi-enzyme systems*. Cambridge University: New York, 1994.

26. Michaelis, L.; Menten, M.L., Die kinetik der invertinwirkung. *Biochem. Zeitschrift* **1913**, *49*, 333.

27. Voit, E.O.; Sands, P.J., Modeling forest growth i. Canonical approach. *Ecol. Model* **1996**, *86*, 51.

28. Savageau, M.A., Biochemical systems analysis. I. Some mathematical properties of the rate law for the component enzymatic reactions. *J Theor Biol* **1969a**, *25*, 365-369.

29. Savageau, M.A., Biochemical systems analysis. Ii. The steady-state solutions for an n-pool system using a power-law approximation. *J Theor Biol* **1969b**, *25*, 370-379.

30. Voit, E.O., *Computational analysis of biochemical systems: A practical guide for biochemists and molecular biologists* Cambridge University: Cambridge, UK, 2000.

31. Kimura, S.; Ide, K.; Kashihara, A.; Kano, M.; Hatakeyama, M.; Masui, R.; Nakagawa, N.; Yokoyama, S.; Kuramitsu, S.; Konagaya, A., Inference of s-system models of genetic networks using a cooperative coevolutionary algorithm. *Bioinformatics* **2005**, *21*, 1154-1163.

32. Atkinson, M.R.; Savageau, M.A.; Myers, J.T.; Ninfa, A.J., Development of genetic circuitry exhibiting toggle switch or oscillatory behavior in escherichia coli. *Cell* **2003**, *113*, 597-607.

33. Vera, J.; Balsa-Canto, E.; Wellstead, P.; Banga, J.R.; Wolkenhauer, O., Power-law models of signal transduction pathways. *Cell Signal* **2007**, *19*, 1531-1541.

34. Savageau, M.A.; Voit, E.O., Recasting nonlinear differential-equations as s-systems - a canonical nonlinear form. *Mathematical Biosciences* **1987**, *87*, 83-115.

35. Voit, E.O., S-system modeling of complex-systems with chaotic input. *Environmetrics* **1993**, *4*, 153-186.

36. Alvarez-Vasquez, F.; Sims, K.J.; Hannun, Y.A.; Voit, E.O., Integration of kinetic information on yeast sphingolipid metabolism in dynamical pathway models. *J Theor Biol* **2004**, *226*, 265-291.

37. Kanehisa, M., The kegg database. *Novartis Foundation Symposium* **2002**, *p*.

38. Kanehisa, M.; Goto, S.; Kawashima, S.; Okuno, Y.; Hattori, M., The kegg resource for deciphering the genome. *Nucleic Acids Res* **2004**, *32*.

39. Caspi, R.; Foerster, H.; Fulcher, C.A.; Hopkinson, R.; Ingraham, J.; Kaipa, P.; Krummenacker, M.; Paley, S.; Pick, J.; Rhee, S.Y., *et al.*, Metacyc: A multiorganism database of metabolic pathways and enzymes. *Nucleic Acids Res* **2006**, *34*.

40. Schomburg, I.; Chang, A.; Ebeling, C.; Gremse, M.; Heldt, C.; Huhn, G.; Schomburg, D., Brenda, the enzyme database: Updates and major new development. *Nucleic Acids Res* **2004**, *32*.

41. Goel, G.; Chou, I.C.; Voit, E.O., Biological systems modeling and analysis: A biomolecular technique of the twenty-first century. *J Biomol Tech* **2006**, *17*, 252-269.

42. Mao, F.; Dam, P.; Wu, H.; Chou, I.-C.; Voit, E.; Xu, Y., *"Prediction of biological pathways through data mining and information fusion", in computational methods for understanding bacterial and archaeal genomes*. Imperial College Press: London, 2008.

43. Schmidt, K.; Nielsen, J.; Villadsen, J., Quantitative analysis of metabolic fluxes in escherichia coli, using two-dimensional nmr spectroscopy and complete isotopomer models. *J Biotechnol* **1999**, *71*, 175-189.

44. Wiechert, W.; Mollney, M.; Petersen, S.; de Graaf, A.A., A universal framework for 13c metabolic flux analysis. *Metab Eng* **2001**, *3*, 265-283.

45. Shirai, T.; Matsuzaki, K.; Kuzumoto, M.; Nagahisa, K.; Furusawa, C.; Shioya, S.; Shimizu, H., Precise metabolic flux analysis of coryneform bacteria by gas chromatography-mass spectrometry and verification by nuclear magnetic resonance. *J Biosci Bioeng* **2006**, *102*, 413-424.

46. Neves, A.R.; Ventura, R.; Mansour, N.; Shearman, C.; Gasson, M.J.; Maycock, C.; Ramos, A.; Santos, H., Is the glycolytic flux in lactococcus lactis primarily controlled by the redox charge? Kinetics of nad(+) and nadh pools determined in vivo by 13c nmr *J. Biol. Chem.* **2002**, *277*, 28088.

47. Szyperski, T., 13c-nmr, ms and metabolic flux balancing in biotechnology research. *Q. Rev. Biophys.* **1998**, *31*.

48. Goodenowe, D., *Metabolomic analysis with fourier transform ion cyclotron resonance mass spectrometry*. Kluwer, Dordrecht, The Netherlands, 2003.

49. Plumb, R.S.; Stumpf, C.L.; Gorenstein, M.V.; Castro-Perez, J.M.; Dear, G.J.; Anthony, M.; Sweatman, B.C.; Connor, S.C.; Haselden, J.N., Metabonomics: The use of electrospray mass spectrometry coupled to reversed-phase liquid chromatography shows potential for the screening of rat urine in drug development, rapid commun. Mass spectrom. **2002**, *16*.

50. Ostergaard, S.; Olsson, L.; Nielsen, J., In vivo dynamics of galactose metabolism in saccharomyces cerevisiae: Metabolic fluxes and metabolite levels. *Biotechnol. Bioeng.* **2001**, *73*.

51. Theobald, U.; Mailinger, W.; Baltes, M.; Rizzi, M.; Reuss, M., In vivo analysis of metabolic dynamics in saccharomyces cerevisiae: I. Experimental observations. *Biotechnol. Bioeng.* **1997**, *55*.

52. Neves, A.R.; Ramos, A.; Nunes, M.C.; Kleerebezem, M.; Hugenholtz, J.; de Vos, W.M.; Almeida, J.; Santos, H., In vivo nuclear magnetic resonance studies of glycolytic kinetics in lactococcus lactis. *Biotechnol Bioeng* **1999**, *64*, 200-212.

53. Famili, I.; Palsson, B.O., The convex basis of the left null space of the stoichiometric matrix leads to the definition of metabolically meaningful pools. *Biophys J* **2003**, *85*, 16-26.

54. De Boor, C., *A practical guide to splines*. Springer-Verlag: New York, 1978; p xxiv, 392 p.

55. De Boor, C.; Höllig, K.; Riemenschneider, S.D., *Box splines*. Springer-Verlag: New York, 1993; p xvii, 200 p.

56. Green, P.J.; Silverman, B.W., *Nonparametric regression and generalized linear models : A roughness penalty approach*. 1st ed.; Chapman & Hall: London ; New York, 1994; p ix, 182 p.

57. Montgomery, D.C.; Runger, G.C., *Applied statistics and probability for engineers*. John Wiley & Sons Pte Ltd: 2007; Vol. 11.

58. Whittaker, E.T., On a new method of graduation. *Proc. Edinburgh Math. Soc.* **1923**, *41*, 63.

59. Almeida, J.S., Predictive non-linear modeling of complex data by artificial neural networks. *Curr Opin Biotechnol* **2002**, *13*, 72-76.

60. Voit, E.O.; Almeida, J., Decoupling dynamical systems for pathway identification from metabolic profiles. *Bioinformatics* **2004**, *20*, 1670-1681.

61. Vilela, M.; Borges, C.C.; Vinga, S.; Vasconcelos, A.T.; Santos, H.; Voit, E.O.; Almeida, J.S., Automated smoother for the numerical decoupling of dynamics models. *BMC bioinformatics* **2007**, *8*, 305.

62. Ljung, L., *System identification : Theory for the user*. 2nd ed.; Prentice Hall: Upper Saddle River, N.J., 1999; p xxii, 609 p.

63. Raue, A.; Kreutz, C.; Maiwald, T.; Bachmann, J.; Schilling, M.; Klingmuller, U.; Timmer, J., Structural and practical identifiability

analysis of partially observed dynamical models by exploiting the profile likelihood. *Bioinformatics* **2009**, *25*, 1923-1929.

64. Srinath, S.; Gunawan, R., Parameter identifiability of power-law biochemical system models. *J Biotechnol* **2010**, *149*, 132-140.

65. Tsai, K.Y.; Wang, F.S., Evolutionary optimization with data collocation for reverse engineering of biological networks. *Bioinformatics* **2005**, *21*, 1180-1188.

66. Voit, E.O., *Computational analysis of biochemical systems : A practical guide for biochemists and molecular biologists*. Cambridge University Press: New York, 2000; p xii, 531 p.

67. Chou, I.C.; Martens, H.; Voit, E.O., Parameter estimation in biochemical systems models with alternating regression. *Theor Biol Med Model* **2006**, *3*, 25.

68. Hlavacek, W.S.; Savageau, M.A., Rules for coupled expression of regulator and effector genes in inducible circuits. *J Mol Biol* **1996**, *255*, 121-139.

69. Wang, F.S.; Su, T.L.; Jang, H.J., Hybrid differential evolution for problems of kinetic parameter estimation and dynamic optimization of an ethanol fermentation process. *Indust. Eng. Chem. Res* **2001**, *40*.

70. Ko, C.L.; Wang, F.S.; Chao, Y.P.; Chen, T.W., S-system approach to modeling recombinant escherichia coli growth by hybri differential evolution with data collocation. *Biochem Eng J* **2006**, *28*, 10-16.

71. Vera, J.; de Atauri, P.; Cascante, M.; Torres, N.V., Multicriteria optimization of biochemical systems by linear programming: Application to production of ethanol by saccharomyces cerevisiae. *Biotechnol Bioeng* **2003**, *83*, 335-343.

72. Curto, R.; Sorribas, A.; Cascante, M., Comparative characterization of the fermentation pathway of saccharomyces cerevisiae using biochemical systems theory and metabolic control analysis: Model definition and nomenclature. *Math Biosci* **1995**, *130*, 25-50.

73. Polisetty, P.K.; Voit, E.O.; Gatzke, E.P., Identification of metabolic system parameters using global optimization methods. *Theoretical biology & medical modelling* **2006**, *3*, 4.

74. Polisetty, P.K.; Gatzke, E.P.; Voit, E.O., Yield optimization of regulated metabolic systems using deterministic branch-and-reduce methods. *Biotechnology and Bioengineering* **2008**, *99*, 1154-1169.

75. Lall, R.; Voit, E.O., Parameter estimation in modulated, unbranched reaction chains within biochemical systems. *Comput Biol Chem* **2005**, *29*, 309-318.

76. Neves, A.R.; Pool, W.A.; Kok, J.; Kuipers, O.P.; Santos, H., Overview on sugar metabolism and its control in lactococcus lactis - the input from in vivo nmr. *Fems Microbiol Rev* **2005**, *29*, 531-554.

77. Goel, G.; Chou, I.C.; Voit, E.O., System estimation from metabolic time-series data. *Bioinformatics* **2008**, *24*, 2505-2511.

78. Voit, E.O.; Almeida, J.; Marino, S.; Lall, R.; Goel, G.; Neves, A.R.; Santos, H., Regulation of glycolysis in lactococcus lactis: An unfinished systems biological case study. *Syst Biol (Stevenage)* **2006**, *153*, 286-298.

79. Chou, I.C.; Voit, E.O., Estimation of dynamic flux profiles from metabolic time series data. *BMC Syst Biol* **2012**, *6*, 84.

80. Fonseca, L.L.; Sanchez, C.; Santos, H.; Voit, E.O., Complex coordination of multi-scale cellular responses to environmental stress. *Mol Biosyst* **2011**, *7*, 731-741.

81. Kimura, S.; Hatakeyama, M.; Konagaya, A., Inference of s-system models of genetic networks from noisy time-series data. *Chem-Bio Inform. J.* **2004**, *4*.

82. Maki, Y.; Tominaga, D.; Okamoto, M.; Watanabe, S.; Eguchi, Y., Development of a system for the inference of large scale genetic networks. *Pac. Symp. Biocomput.* **2001**, *446*.

83. Tucker, W.; Kutalik, Z.; Moulton, V., Estimating parameters for generalized mass action models using constraint propagation. *Math. Biosci.* **2007**, *208*, 607.

84. Sands, P.J.; Voit, E.O., Flux-based estimation of parameters in s-systems. *Ecol. Model* **1996**, *93*.

85. Bardow, A.; Marquardt, W., Incremental and simultaneous identification of reaction kinetics: Methods and comparison. *Chem Eng Sci* **2004**, *59*, 2673-2684.

86. Voit, E.O., Symmetries of s-systems. *Math. Biosci.* **1992**, *109*.

87. Voit, E.O.; Goel, G.; Chou, I.C.; Fonseca, L.L., Estimation of metabolic pathway systems from different data sources. *IET Syst Biol* **2009**, *3*, 513-522.

88. Bonvin, D.; Rippin, D.W.T., Target factor-analysis for the identification of stoichiometric models. *Chem Eng Sci* **1990**, *45*, 3417-3426.

89. Chevalier, T.; Schreiber, I.; Ross, J., Toward a systematic determination of complex-reaction mechanisms. *J Phys Chem-Us* **1993**, *97*, 6776-6787.

90. Arkin, A.; Ross, J., Statistical construction of chemical-reaction mechanisms from measured time-series. *J Phys Chem-Us* **1995**, *99*, 970-979.

91. Tominaga, D.; Koga, N.; Okamoto, M., Efficient numerical optimization algorithm based on genetic algorithm for inverse problem. *Proceedings of the Genetic and Evolutionary Computation Conference* **2000**, 251.

92. Samoilov, M.; Arkin, A.; Ross, J., On the deduction of chemical reaction pathways from measurements of time series of concentrations. *Chaos* **2001**, *11*, 108-114.

93. Maki, Y.; Ueda, T.; Masahiro, O.; Naoya, U.; Kentaro, I.; Uchida, K., Inference of genetic network using the expression profile time course data of mouse p19 cells. *Genome Informatics* **2002**, *13*, 382-383.

94. Vance, W.; Arkin, A.; Ross, J., Determination of causal connectivities of species in reaction networks. *P Natl Acad Sci USA* **2002**, *99*, 5816-5821.

95. Kikuchi, S.; Tominaga, D.; Arita, M.; Takahashi, K.; Tomita, M., Dynamic modeling of genetic networks using genetic algorithm and s-system. *Bioinformatics* **2003**, *19*, 643-650.

96. Veflingstad, S.R.; Almeida, J.; Voit, E.O., Priming nonlinear searches for pathway identification. *Theor Biol Med Model* **2004**, *1*, 8.

97. Crampin, E.J.; McSharry, P.E.; Schnell, S., Extracting biochemical reaction kinetics from time series data. *Knowledge-Based Intelligent Information and Engineering Systems, Pt 2, Proceedings* **2004**, *3214*, 329-336.

98. Katare, S.; Kalos, A.; West, D., A hybrid swarm optimizer for efficient parameter estimation. *Cec2004: Proceedings of the 2004 Congress on Evolutionary Computation, Vols 1 and 2* **2004**, 309-315.

99. Kimura, S.; Ide, K.; Kashihara, A.; Kano, M.; Hatakeyama, M.; Masui, R.; Nakagawa, N.; Yokoyama, S.; Kuramitsu, S.; Konagaya, A., Inference of s-system models of genetic networks using a cooperative coevolutionary algorithm. *Bioinformatics* **2005**, *21*, 1154-1163.

100. Tucker, W.; Moulton, V., Reconstructing metabolic networks using interval analysis. *Lect Notes Comput Sc* **2005**, *3692*, 192-203.

101. Marino, S.; Voit, E.O., An automated procedure for the extraction of metabolic network information from time series data. *Journal of bioinformatics and computational biology* **2006**, *4*, 665-691.

102. Marquardt, W.; Brendel, M.; Bonvin, D., Incremental identification of kinetic models for homogeneous reaction systems. *Chem Eng Sci* **2006**, *61*, 5404-5420.

103. Daisuke, T.; Horton, P., Inference of scale-free networks from gene expression time series. *J Bioinform Comput Biol* **2006**, *4*, 503-514.

104. Cho, D.Y.; Cho, K.H.; Zhang, B.T., Identification of biochemical networks by s-tree based genetic programming. *Bioinformatics* **2006**, *22*, 1631-1640.

105. Kutalik, Z.; Tucker, W.; Moulton, V., S-system parameter estimation for noisy metabolic profiles using newton-flow analysis. *IET Syst Biol* **2007**, *1*, 174-180.

106. Tucker, W.; Kutalik, Z.; Moulton, V., Estimating parameters for generalized mass action models using constraint propagation. *Math Biosci* **2007**, *208*, 607-620.

107. Noman, N.; Iba, H., Inferring gene regulatory networks using differential evolution with local search heuristics. *IEEE/ACM transactions on computational biology and bioinformatics / IEEE, ACM* **2007**, *4*, 634-647.

108. Gonzalez, O.R.; Kuper, C.; Jung, K.; Naval, P.C., Jr.; Mendoza, E., Parameter estimation using simulated annealing for s-system models of biochemical networks. *Bioinformatics* **2007**, *23*, 480-486.

109. Vilela, M.; Chou, I.C.; Vinga, S.; Vasconcelos, A.T.; Voit, E.O.; Almeida, J.S., Parameter optimization in s-system models. *BMC Syst Biol* **2008**, *2*, 35.

110. Zuniga, P.C.; Pasia, J.; Adorna, H.; del Rosario, R.C.H.; Naval, P., An ant colony optimization algorithm for parameter estimation and network inference problems in s-system models. *International Conference on Molecular Systems Biology 2008* **2008**, 105.

111. Machina, A.; Ponosov, A.; Voit, E.O., Automated piecewise power-law modeling of biological systems. *J Biotechnol* **2010**, *149*, 154-165.

112. Zhan, C.J.; Yeung, L.F., Parameter estimation in systems biology models using spline approximation. *Bmc Systems Biology* **2011**, *5*.

113. Chevalier, T.; Schreiber, I.; Ross, J., Toward a systematic determination of complex reaction mechanisms. *J. Phys. Chem.* **1993**, *97*, 6776.

114. Vance, W.; Arkin, A.; Ross, J., Determination of causal connectivities of species in reaction networks. *Proc Natl Acad Sci U S A* **2002**, *99*, 5816-5821.

115. Arkin, A.; Ross, J., Statistical construction of chemical-reaction mechanisms from measured time-series. *J. Phys. Chem.* **1995**, *99*.

116. Arkin, A.; Shen, P.D.; Ross, J., A test case of correlation metric construction of a reaction pathway from measurements. *Science* **1997**, *277*.

117. Amrhein, M.; Srinivasan, B.; Bonvin, D., Target factor analysis of reaction data: Use of data pre-treatment and reaction-invariant relationships. *Chemical Engineering Science* **1999**, *54*, 579-591.

118. Bonvin, D.; Rippin, D.W.T., Target factor analysis for the identification of stoichiometric models. *Chemical Engineering Science* **1990**, *45*, 3417-3426.

119. Hendry, D.F.; Krolzig, H.M., *New developments in automatic general-to-specific modelling*. Princeton University: 2003; p 379-419.

120. Crampin, E.J.; Schnell, S.; McSharry, P.E., Mathematical and computational techniques to deduce complex biochemical reaction mechanisms. *Prog Biophys Mol Bio* **2004**, *86*, 77-112.

121. Thieffry, D.; Huerta, A.M.; Perez-Rueda, E.; Collado-Vides, J., From specific gene regulation to genomic networks: A global analysis of transcriptional regulation in escherichia coli. *Bioessays* **1998**, *20*, 433-440.

122. Jeong, H.; Tombor, B.; Albert, R.; Oltvai, Z.N.; Barabasi, A.L., The large-scale organization of metabolic networks. *Nature* **2000**, *407*, 651-654.

123. Noman, N.; Iba, H., Reverse engineering genetic networks using evolutionary computation. *Genome Inform* **2005**, *16*, 205-214.

124. Edwards, J.S.; Covert, M.; Palsson, B., Metabolic modelling of microbes: The flux-balance approach. *Environ Microbiol* **2002**, *4*, 133-140.

125. Chassagnole, C.; Noisommit-Rizzi, N.; Schmid, J.W.; Mauch, K.; Reuss, M., Dynamic modeling of the central carbon metabolism of escherichia coli. *Biotechnology and Bioengineering* **2002**, *79*, 53-73.

126. Curto, R.; Voit, E.O.; Sorribas, A.; Cascante, M., Mathematical models of purine metabolism in man. *Mathematical Biosciences* **1998**, *151*, 1-49.

127. Marquardt, D.W., An algorithm for least-squares estimation of nonlinear parameters. *J Soc Ind Appl Math* **1963**, *11*, 431-441.

128. Mendes, P.; Kell, D., Non-linear optimization of biochemical pathways: Applications to metabolic engineering and parameter estimation. *Bioinformatics* **1998**, *14*, 869-883.

129. Land, A.H.; Doig, A.G., An automatic method of solving discrete programming-problems. *Econometrica* **1960**, *28*, 497-520.

130. Ashyraliyev, M.; Fomekong-Nanfack, Y.; Kaandorp, J.A.; Blom, J.G., Systems biology: Parameter estimation for biochemical models. *Febs Journal* **2009**, *276*, 886-902.

131. Goldberg, D.E., *Genetic algorithms in search, optimization and machine learning*. Addison-Wesley Pub. Co.: Reading, Mass. ; Wokingham, 1989.

132. Okamoto, M.; Nonaka, T.; Ochiai, S.; Tominaga, D., Nonlinear numerical optimization with use of a hybrid genetic algorithm incorporating the modified powell method. *Appl Math Comput* **1998**, *91*, 63-72.

133. Ueda, T.; Koga, N.; Okamoto, M., Efficient numerical optimization technique based on real-coded genetic algorithm. *Genome Informatics* **2001**, *12*, 451–453.

134. Ueda, T.; Ono, I.; Okamoto, M., Development of system identification technique based on real-coded genetic algorithm. *Genome Informatics* **2002**, *13*, 386-387.

135. Ho, S.Y.; Hsieh, C.H.; Yu, F.C.; Huang, H.L., An intelligent two-stage evolutionary algorithm for dynamic pathway identification from gene expression profiles. *IEEE/ACM Trans Comput Biol Bioinform* **2007**, *4*, 648-660.

136. Storn, R.; Price, K., Differential evolution - a simple and efficient heuristic for global optimization over continuous spaces. *J Global Optim* **1997**, *11*, 341-359.

137. Storn, R., On the usage of differential evolution for function optimization. *1996 Biennial Conference of the North American Fuzzy Information Processing Society - Nafips* **1996**, 519-523.

138. Noman, N.; Iba, H., Inference of genetic networks using s-system: Information criteria for model selection. *Gecco 2006: Genetic and Evolutionary Computation Conference, Vol 1 and 2* **2006**, 263-270.

139. Noman, N.; Iba, H., Inferring gene regulatory networks using differential evolution with local search heuristics. *Ieee Acm T Comput Bi* **2007**, *4*, 634-647.

140. Noman, N.; Iba, H., Inference of gene regulatory networks using s-system and differential evolution. *Gecco 2005: Genetic and Evolutionary Computation Conference, Vols 1 and 2* **2005**, 439-446.

141. Liu, P.K.; Wang, F.S., Inference of biochemical network models in s-system using multiobjective optimization approach. *Bioinformatics* **2008**, *24*, 1085-1092.

142. Koza, J.R., *Genetic programming : On the programming of computers by means of natural selection*. MIT Press: Cambridge, Mass., 1992; p xiv, 819 p.

143. Koza, J.R.; Mydlowec, W.; Lanza, G.; Yu, J.; Keane, M.A., Reverse engineering of metabolic pathways from observed data using genetic programming. *Pac Symp Biocomput* **2001**, 434-445.

144. Sugimoto, M.; Kikuchi, S.; Tomita, M., Reverse engineering of biochemical equations from time-course data by means of genetic programming. *Biosystems* **2005**, *80*, 155-164.

145. Moles, C.G.; Mendes, P.; Banga, J.R., Parameter estimation in biochemical pathways: A comparison of global optimization methods. *Genome Res* **2003**, *13*, 2467-2474.

146. Spieth, C.; Worzischek, R.; Streichert, F., Comparing evolutionary algorithms on the problem of network inference. *Gecco 2006: Genetic and Evolutionary Computation Conference, Vol 1 and 2* **2006**, 305-306.

147. Dorigo, M.; Stützle, T., *Ant colony optimization*. MIT Press: Cambridge, Mass., 2004; p xi, 305 p.

148. Clerc, M., *Particle swarm optimization*. ISTE: London, 2006; p 243 p.

149. Naval, P.C.; Sison, L.G.; Mendoza, E., Metabolic network parameter inference using particle swarm optimization. *International Conference on Molecular Systems Biology 2006* **2006**.

150. Kirkpatrick, S.; Gelatt, C.D.; Vecchi, M.P., Optimization by simulated annealing. *Science* **1983**, *220*, 671-680.

151. Ceric, S.; Kurtanjek, Z., Model identification, parameter estimation, and dynamic flux analysis of e-coli central metabolism. *Chem Biochem Eng Q* **2006**, *20*, 243-253.

152. Katare, S.; Bhan, A.; Caruthers, J.M.; Delgass, W.N.; Venkatasubramanian, V., A hybrid genetic algorithm for efficient parameter estimation of large kinetic models. *Comput Chem Eng* **2004**, *28*, 2569-2581.

153. Runarsson, T.P.; Yao, X., Stochastic ranking for constrained evolutionary optimization. *Ieee T Evolut Comput* **2000**, *4*, 284-294.

154. Rodriguez-Fernandez, M.; Mendes, P.; Banga, J.R., A hybrid approach for efficient and robust parameter estimation in biochemical pathways. *Biosystems* **2006**, *83*, 248-265.

155. Fomekong-Nanfack, Y.; Kaandorp, J.A.; Blom, J., Efficient parameter estimation for spatio-temporal models of pattern formation: Case study of drosophila melanogaster. *Bioinformatics* **2007**, *23*, 3356-3363.

156. Ashyraliyev, M.; Jaeger, J.; Blom, J.G., Parameter estimation and determinability analysis applied to drosophila gap gene circuits. *Bmc Systems Biology* **2008**, *2*.

157. Egea, J.A.; Rodriguez-Fernandez, M.; Banga, J.R.; Marti, R., Scatter search for chemical and bio-process optimization. *J Global Optim* **2007**, *37*, 481-503.

158. Rodriguez-Fernandez, M.; Egea, J.A.; Banga, J.R., Novel metaheuristic for parameter estimation in nonlinear dynamic biological systems. *BMC bioinformatics* **2006**, *7*.

159. Laguna, M.; Marti, R., Experimental testing of advanced scatter search designs for global optimization of multimodal functions. *J Global Optim* **2005**, *33*, 235-255.

160. Vilela, M.; Vinga, S.; Maia, M.A.; Voit, E.O.; Almeida, J.S., Identification of neutral biochemical network models from time series data. *BMC Syst Biol* **2009**, *3*, 47.

161. Montgomery, D.C.; Runger, G.C., *Applied statistics and probability for engineers*. 4th ed.; Wiley: Hoboken, NJ, 2007; p xvi, 768 p.

162. Yao, K.Z.; Shaw, B.M.; Kou, B.; McAuley, K.B.; Bacon, D.W., Modeling ethylene/butene copolymerization with multi-site catalysts: Parameter estimability and experimental design. *Polym React Eng* **2003**, *11*, 563-588.

163. Ramos, A.; Neves, A.R.; Santos, H., Metabolism of lactic acid bacteria studied by nuclear magnetic resonance. *Anton Leeuw Int J G* **2002**, *82*, 249-261.

164. Liu, P.K.; Wang, F.S., Inverse problems of biological systems using multi-objective optimization. *J Chin Inst Chem Eng* **2008**, *39*, 399-406.

165. Gennemark, P.; Wedelin, D., Efficient algorithms for ordinary differential equation model identification of biological systems. *Iet Systems Biology* **2007**, *1*, 120-129.

166. Ramsay, J.O.; Hooker, G.; Campbell, D.; Cao, J., Parameter estimation for differential equations: A generalized smoothing approach. *J Roy Stat Soc B* **2007**, *69*, 741-770.

167. Akaike, H., New look at statistical-model identification. *Ieee T Automat Contr* **1974**, *Ac19*, 716-723.

168. Neves, A.R.; Ramos, A.; Costa, H.; van, S., II; Hugenholtz, J.; Kleerebezem, M.; de Vos, W.; Santos, H., Effect of different nadh oxidase levels on glucose metabolism by lactococcus lactis: Kinetics of intracellular metabolite pools determined by in vivo nuclear magnetic resonance. *Appl Environ Microbiol* **2002**, *68*, 6332-6342.

169. Eilers, P.H., A perfect smoother. *Anal Chem* **2003**, *75*, 3631-3636.

170. Jaulin, L.; Kieffer, M.; Didrit, O.; Walter, E., *Applied interval analysis : With examples in parameter and state estimation, robust control and robotics*. Springer: London, 2001; p xvi, 379 p.

171. Lin, Y.D.; Stadtherr, M.A., Validated solution of odes with parametric uncertainties. *Comput-Aided Chem En* **2006**, *21*, 167-172.

172. Imoto, S.; Kim, S.; Goto, T.; Miyano, S.; Aburatani, S.; Tashiro, K.; Kuhara, S., Bayesian network and nonparametric heteroscedastic regression for nonlinear modeling of genetic network. *J Bioinform Comput Biol* **2003**, *1*, 231-252.

173. Nagarajan, R.; Upreti, M., Comment on causality and pathway search in microarray time series experiment. *Bioinformatics* **2008**, *24*, 1029-1032.

174. Tung, T.Q.; Ryu, T.; Lee, K.H.; Lee, D., Inferring gene regulatory networks from microarray time series data using transfer entropy. *Comp Med Sy* **2007**, 383-388.

175. Latendresse, M.; Paley, S.; Karp, P.D., Browsing metabolic and regulatory networks with biocyc. *Methods Mol Biol* **2012**, *804*, 197-216.

176. Zamora-Sillero, E.; Hafner, M.; Ibig, A.; Stelling, J.; Wagner, A., Efficient characterization of high-dimensional parameter spaces for systems biology. *Bmc Systems Biology* **2011**, *5*.

177. Newman, M.E.J.; Barkema, G.T., *Monte carlo methods in statistical physics*. Clarendon Press: Oxford, 1999; p xiv, 475 p.

178. Khachiyan, L.G., Rounding of polytopes in the real number model of computation. *Math Oper Res* **1996**, *21*, 307-320.

179. Bard, Y., *Nonlinear parameter estimation*. Academic Press: New York London, 1974; p x,341p.

180. Miskovic, L.; Hatzimanikatis, V., Modeling of uncertainties in biochemical reactions. *Biotechnology and Bioengineering* **2011**, *108*, 413-423.

181. Nielsen, J., Metabolic engineering: Techniques for analysis of targets for genetic manipulations. *Biotechnol Bioeng* **1998**, *58*, 125-132.

182. Gunawan, R., Framework for the creation and applications of ensemble modeling in systems biology and biotechnology. **2012**.

183. Chih-Lung, K.; Feng-Sheng, W.; Yun-Peng, C.; Te-Wei, C., S-system approach to modeling recombinant escherichia coli growth by hybrid differential evolution with data collocation. *Bochemical Engineering J.* **2006**, *28*.

APPENDIX A

A1 A Generic Branched Pathway

Table A1 summarizes the parameter values of this generic branched pathway, including their true values. As a complete comparison, ODE decomposition and two-phase estimation methods were both applied to the cases under the following three conditions: noise-free data with X_3 missing, noisy data with X_3 missing and noise-free data with X_2 X_3 both missing (Figure A1). Given half information (X_2 X_3 both missing), it was expected that all the indexes would increase in Table A2, but the proposed method was still better than the ODE decomposition alone, in terms of reduced slope error and concentration error at more than half reduced computational cost. In addition, parameter estimates from both methods were able to capture the trend of X_2, but the proposed method can also follow the rough trend of X_3.

Table A1. Parameter values in the branched metabolic pathway model

	Is the parameter *a priori* identifiable ?[a]	**True Values** [60]	**ODE Decomposition**			**Two-Phase Estimation**		
			w/o noise (X_3 missing)	**w/ noise (X_3 missing)**	**w/o noise (X_2 and X_3 missing)**	**w/o noise (X_3 missing)**	**w noise (X_3 missing)**	**w/o noise (X_2 and X_3 missing)**
α_1	N	20	14.3251	7.4638	8.6907	7.3915	11.9498	0.3621
β_1	Y	10	15.5590	7.8476	20.0081	7.1857	9.4370	8.0324
α_2	N	8	7.1566	7.6614	18.3646	7.0850	9.1653	0
β_2	Y	3	2.1827	3.1131	24.4023	2.6401	5.2429	1.9480
β_3	Y	5	21.2695	14.8967	24.7699	6.3065	9.9220	5.5003
β_4	N	6	21.6201	2.6303	10.7145	4.4590	2.6462	9.7779
g_{13}	Y	-0.8	-0.7376	-1.8268	1.9996	-0.3467	-0.2569	2.0000
h_{11}	Y	0.5	0.6008	0.5623	0.9972	0.5065	0.2896	0.0625
h_{22}	N	0.75	0.9141	0.5854	0.1257	0.7601	0.4338	2.0000
h_{33}	N	0.5	0.6717	2.0000	1.2316	0.2599	0.2288	-0.4420
h_{34}	N	0.2	0.8234	0.5852	0.7596	0.1676	0.1727	2.0000
h_{44}	N	0.8	0.7780	1.5199	1.4350	2.0000	1.2618	0.1298
$X_3(t_0)$	N	1.2	0.3879	0.7886	2.3493	0.2216	0.7612	5
$X_2(t_0)$	N	2.7	—	—	0.7374	—	—	1.325

[a] *A priori* identifiable parameter (AIP) with missing X_3 data. The *a priori* identifiability was determined using orthogonal decomposition of the sensitivity matrix [162].

Table A2. Parameter estimation of the branched pathway model

	ODE Decomposition			**Two-Phase Estimation**		
	w/o noise (X_3 missing)	**w/ noise (X_3 missing)**	**w/o noise (X_2 and X_3 missing)**	**w/o noise (X_3 missing)**	**w/ noise (X_3 missing)**	**w/o noise (X_2 and X_3 missing)**
Computational time (sec) [a]	4493.2	10910.3	180045	1062.1	2807.4	88667.4
Number of stiff ODE simulations	1247	2012	9173	359	823	4401
Parameter error	92.18%	90.97%	209.31%	36.59%	47.27%	175.00%
Slope error [b]	2.5962	9.4303	2.6321	0.8620	8.5909	2.5389
Concentration error [c]	0.5137	5.8207	0.0533	0.1526	3.6021	0.0186

[a] The computational time was based on Dual Processors Intel Quad-Core 2.83 GHz.

[b] Slope error was calculated using Equation 3.3, in which $\mathbf{X}_u$, $\mathbf{X}_m$ are from simultaneous ODE simulation.

[c] Concentration error was calculated using Equation 3.4, in which $\mathbf{X}_m$ are from simultaneous ODE simulation.

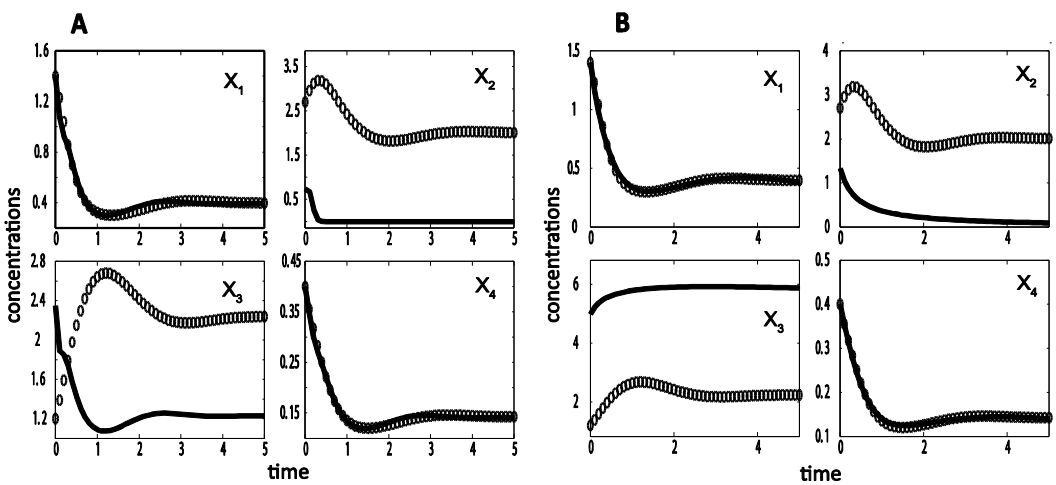


Figure A1. ODE decomposition parameter estimation (**A**) and two-phase estimation (**B**) in the branched pathway model: concentration simulations for the case where both X_2 and X_3 are missing; (—) simulation profile, (○) *in silico* data.

A2 *E. coli* Metabolism Model

Table A3 presents the parameter values and initial concentration of X_2 in an *E. coli* model [70]. Data in the first column are the values reported by Chih-Lung et al., and the parameter values of the second column are estimated based on complete data using decoupling method. The true values for $X_{20(1)}$, $X_{20(2)}$ are directly obtained from the data by taking average on the duplicates of the initial glucose concentrations. The third and fourth columns contain the estimates from ODE decomposition method and the proposed method respectively, given incomplete experimental data (measurements of X_2 are completely missing).

Table A3. Parameter values in the *E. coli* model

	Parameter estimates from complete data		**Parameter estimates from incomplete data**	
	Previous report [183]	**Decoupling method**	**ODE decomposition**	**Proposed method**
α_1	0.1891	0.0088	0.3883	0.0010
β_2	0.6917	1.0448	1.8627	0.1969
α_3	0.0655	0.0026	0.2182	0.0010
α_4	1.2010	0.4513	1.9847	0.0010
α_5	0.2493	0.2470	0.1889	0.2460
g_{11}	0.0100	0.5980	0.2762	0.4682
g_{12}	0.2118	1.0505	0.1989	1.4741
h_{21}	1.7219	0.9059	1.7259	0.9941
h_{22}	0.2126	0.2793	1.3655	0.6279

g_{31}	0.0100	0.5160	0.4642	0.3341
g_{32}	0.3033	1.1891	0.1976	1.2915
g_{41}	0.8578	0.9907	0.8242	1.4235
g_{42}	0.1080	0.4014	0.1237	1.8239
g_{51}	0.0497	0.1887	0.2233	0.1890
g_{52}	0.0100	0	0.0902	0.0010
Initial concentration of X_2	**True values**		**Estimated values**	
$X_2(t_0)$	38.933		17.1720	65.1193
$X_2(t_0)'$	49.965		17.5776	65.1193

A3 Glycolytic Pathway in *Lactococcus lactis*

Table A4 summarizes the parameter values of this *L. lactis* metabolic model with ODE decomposition and two-phase methods, given *in silico* data or filtered experimental data with X_3 missing.

Table A4. Parameter values in the *L. lactis* metabolic model

	Parameter values from previous report [160]	**ODE Decomposition**		**Two-Phase Estimation**	
		w/o noise (X_3 missing)	**filtered data (X_3 missing)**	**w/o noise (X_3 missing)**	**filtered data (X_3 missing)**
α_1	1.3113	6.8442	12.1396	2.1655	0.2126
β_1	4.0821	10.279	15.0059	5.0486	2.4090
h_{11}	0.1230	0.0453	0.0303	0.1345	0.2165
h_{14}	0.4142	0.1651	0.0732	0.2665	0.3976

α_2	0.5071	10.5324	3.3458	0.0538	0.7470
g_{21}	0.8844	0.8008	0.8325	1.5938	0.3164
g_{24}	0.1118	0.0243	0.0614	0.1112	0
β_2	0.9852	11.4034	6.8004	0.6320	1.1690
h_{22}	1.0720	1.4443	1.1673	1.1857	0.4771
α_3	12.7563	9.1829	16.0482	5.0910	16.0907
g_{32}	0.7635	0.1634	0.7812	0.2595	1.9114
β_3	7.2386	10.9689	12.2713	4.0641	19.9999
h_{33}	0.3976	0.6603	1.7885	0.1708	1.1560
α_4	5.3176	16.2099	10.2630	0.3023	2.9194
g_{43}	0.1466	0.3471	1.9885	1.0195	0.2588
β_4	6.2504	8.0156	3.4196	0.3563	0.5300
h_{42}	0.3704	0.7773	1.9038	1.4371	1.8527
h_{44}	0.1102	0.9822	1.0493	0.5654	0.2042
α_5	13.8804	6.5624	3.4473	20.0	17.6495
g_{54}	0.2255	0.5162	1.8048	0.1453	0.1383
β_5	8.5617	2.0799	0.0313	14.5981	12.6867
α_6	0.4206	0.4164	0.5316	0.4442	0.4697
g_{64}	0.7670	0.7504	1.8335	0.6177	0.4852
$X_3(t_0)$	0.4000	0.4253	—	0.3467	—
$X_3^*(t_0)$ [a]	9.7381	—	2.3708	—	2.2187

[a] The initial concentration which was used in filtered data

APPENDIX B

Table B1. Parameter estimations of the branched pathway model using noise-free data and analytical slope values

	Simultaneous method		**Incremental method**	
	$\min \Phi_C$ [b]	$\min \Phi_S$ [c]	$\min \Phi_C$ [c]	$\min \Phi_S$ [c]
Computational time (sec) [a]	56.00 hours	785.79 ±50.80	108.91 ±2.99	3.17 $\pm 4.72\times10^{-2}$
Average parameter error	49.10%	4.93×10^{-3}% $\pm 2.84\times10^{-4}$%	5.06×10^{-5}% $\pm 8.47\times10^{-7}$%	1.96×10^{-5}% $\pm 7.63\times10^{-7}$%
$\sqrt{\Phi_C}$ [d]	$\underline{4.54\times10^{-3}}$	2.17×10^{-6} $\pm 4.00\times10^{-9}$	$\underline{5.37\times10^{-9}}$ $\pm 7.52\times10^{-12}$	6.75×10^{-9} $\pm 1.19\times10^{-10}$
$\sqrt{\Phi_S}$ [d]	4.84×10^{-2}	$\underline{3.01\times10^{-6}}$ $\pm 2.57\times10^{-9}$	3.41×10^{-8} $\pm 7.22\times10^{-10}$	$\underline{1.36\times10^{-8}}$ $\pm 6.17\times10^{-11}$

a. The computational time was based on a workstation with dual Intel Quad-Core 2.83 GHz processors.

b. Only one out of five runs was stopped with relative improvement of the objective function below 1% between iterations. The rest did not converge within the 5-day time limit after iterating for 583, 989, 777, and 661 times. The corresponding Φ_C at termination were 4.85×10^{-2}, 1.39×10^{-2}, 1.75×10^{-2} and 3.75×10^{-2}, respectively.

c. Mean value and standard deviation (±) out of five runs, which converged with relative improvement of the objective function below 0.01%.

d. Root mean square error of model predictions and the underlined part refers to the objective function of the minimization.

Table B2. Parameter estimates of the branched metabolic pathway model (simultaneous method)

*	True Values [60]	Simultaneous method $\min \Phi_S$			
		(1)	**(2)**	**(3)**	**(4)**
γ_1	20	19.9999	22.0151	23.2163	20.9692
f_{13}	0.8	0.8000	0.6179	0.2340	0.3713
γ_2	8	7.9998	10.1122	6.5743	9.8968
f_{21}	0.5	0.5000	0.3498	0.5569	0.3599
γ_3	3	2.9998	5.1168	2.3392	4.9036
f_{32}	0.75	0.7500	0.5174	0.7568	0.5342
γ_4	5	4.9998	7.0401	2.7497	9.3560
f_{43}	0.5	0.5000	0.3262	0.4526	0.3054
f_{44}	0.2	0.2000	0.1135	0.0031	0.2082
γ_5	2	2.0002	1.5302	7.6821	4.2064
f_{51}	0.5	0.4999	0.8258	0.0003	0.1642
γ_6	6	5.9997	7.7990	8.4180	6.4270
f_{64}	0.8	0.7999	1.2250	0.0452	0.2945
$X_3(t_0)$	1.2	—	—	—	0.7548

* This table reports the parameter estimates with the minimal objective function value out of five runs.

(1) using noise-free data and analytical slopes; (2) using noise-free data; (3) using noisy data; (4) using noise-free data with missing X_3

Table B3. Parameter estimates of the branched metabolic pathway model (incremental method)

*	True Values [60]	Incremental method $\min \Phi_C$				Incremental method $\min \Phi_S$			
		(1)	**(2)**	**(3)**	**(4)**	**(1)**	**(2)**	**(3)**	**(4)**
γ_1	20	20.0000	20.0105	24.9989	13.4674	20.0000	22.5904	15.0593	24.9585
f_{13}	0.8	0.8000	0.7634	0.3366	1.0920	0.8000	0.6058	0.7824	0.4894
γ_2	8	8.0000	8.7730	14.1896	7.4143	8.0000	10.3220	7.2424	10.1723
f_{21}	0.5	0.5000	0.4410	0.2610	0.5301	0.5000	0.3417	0.4804	0.3479
γ_3	3	3.0000	3.6749	8.6709	2.5980	3.0000	5.2978	2.8968	5.1604
f_{32}	0.75	0.7500	0.6680	0.3577	0.8098	0.7500	0.5072	0.6827	0.5160
γ_4	5	5.0000	5.9268	10.9451	8.2781	5.0000	7.2630	3.4761	7.0669
f_{43}	0.5	0.5000	0.4021	0.1585	0.8642	0.5000	0.3213	0.4371	0.3023
f_{44}	0.2	0.2000	0.1719	0.0579	0.4950	0.2000	0.1133	0.0338	0.1042
γ_5	2	2.0000	1.3828	0.3694	1.6768	2.0000	1.6284	0.8468	3.2351
f_{51}	0.5	0.5000	0.8068	0.0000	1.2353	0.5000	0.7753	1.4665	0.2243
γ_6	6	5.9999	7.3216	1.4041	15.0425	6.0000	7.7068	11.1042	5.7002
f_{64}	0.8	0.8000	1.2352	0.6459	1.7137	0.8000	1.1649	2.0000	0.3960
$X_3(t_0)$	1.2	—	—	—	0.7865	—	—	—	1.2773

* This table reports the parameter estimates with the minimal objective function value out of five runs.

(1) using noise-free data and analytical slopes; (2) using noise-free data; (3) using noisy data; (4) using noise-free data with missing X_3

Table B4. Parameter estimates of the *L. lactis* metabolic model

*	**Simultaneous method**	**Incremental method**	
	$\min \Phi_S$	$\min \Phi_C$	$\min \Phi_S$
γ_1	2.2638	9.7891	0.4994
$f_{1,\,Glu}$	0.0690	0.2627	0.8716
f_{11}	1.2991	0.0309	-1.0343
f_{14}	-0.5461	0.3979	0.9642
γ_2	0.2330	49.9072	49.9999
f_{21}	1.9573	0.4358	0.4404
$f_{2,\,ATP}$	0.9219	-0.3360	-0.8733
γ_3	5.8716	8.3470	5.9069
f_{32}	0.2739	0.4571	0.3602
$f_{3,\,Pi}$	-0.1315	0.1254	0.0477
γ_4	1.5800×10^{-13}	49.6053	0.4193
f_{44}	8.9194×10^{-6}	4.9730	1.7635
γ_5	49.9999	5.2494	49.9999
f_{53}	-0.4609	3.4524	-0.0887
γ_6	3.3189	11.0241	8.2447
f_{62}	0.4006	0.3926	0.2874
f_{64}	0.1383	0.0208	0.2041
$f_{6,\,Pi}$	-0.2920	0.0279	-0.2545
γ_7	0.0001	0	3.0295×10^{-5}
f_{74}	4.9999	1.0855×10^{-7}	0.0005
γ_8	6.3648×10^{-9}	0.5332	0.5332
f_{85}	1.7507	0.1781	0.1781
f_{82}	4.4842	0.4804	0.4804
γ_9	5.4359	34.4010	17.7804

f_{95}	0.5957	0.4394	0.3410

* This table reports the parameter estimates with the minimal objective function value out of five runs.

APPENDIX C

Table C1 summarizes the parameter estimation results in the generation of the initial parameter point for the OEAMC algorithm, for the generic branched pathway example in Chapter 5. The same estimation was repeated for 100 randomly generated data using the same assumption and procedure as done in the case study in Chapter 5. The upper confidence bound for $\sqrt{\Phi_S}$ was estimated to be 2.952×10^{-1}.

Table C1. Parameter estimation of the branched pathway model using Φ_S.

$\sqrt{\Phi_S}$ [a]	1.369×10^{-1}
$\sqrt{\Phi_R}$ [b]	1.380×10^{-1}
$\sqrt{\Phi_C}$ [c]	4.632×10^{-2}

a. Slope error, the minimized objective, was defined by Equation 4.5.
b. Regression error was calculated by Equation 5.2.
c. Concentration error was calculated by Equation 4.4.

Table C2 provides the summary of the ensemble construction based on the slope error function $\sqrt{\Phi_S}$. The volume of the viable subspace of $\mathbf{p}_I$ was 0.2701% of the volume set by the parameter bounds. The range of values for the slope and concentration errors were again computed from uniformly sampling parameter points from the viable space (n = 75680). Figure C1 shows two-dimensional

projections of the viable parameter space onto the parameter axes of fluxes v_1 and v_6. Lastly, Figure C2 compares the metabolite concentration predictions produced by five randomly picked member models and the *in silico* generated noisy data used for the construction of the model ensemble. Again, these models could provide similar goodness-of-fit to the data.

Table C2. Ensemble kinetic modeling of the branched pathway model using Φ_S.

Computational time (sec) [a]	1865
Calculated volume of initial parameter space (V_{ci}) [b]	2.5×10^5
Estimated volume of viable parameter space (V_{ev}) [c]	675.3 ± 4.2
Ratio of V_{ev} to V_{ci}	$(270.1 \pm 1.7) \times10^{-3}$ %
Value range of concentration errors $\sqrt{\Phi_C}$ [d]	$[3.526\times10^{-2}, 2.366\times10^{-1}]$
Value range of slope errors $\sqrt{\Phi_S}$ [e]	$[1.370\times10^{-1}, 2.952\times10^{-1}]$

a. The computational time was the total time of ensemble construction including OEAMC and MEBS phases, based on Dual Processors Intel Quad-Core 2.83 GHz.

b. V_{ci} was calculated through multiplication of initial parameter search ranges (i.e., 100×5×100×5).

c. V_{ev} was calculated by integrating the volume of an ensemble of ellipsoids that cover the viable parameter space [176].

d. Concentration errors were calculated by Equation 4.4, given the parameter samples within the viable parameter space.

e. Slope errors were calculated by Equation 4.5, given the parameter samples within the viable parameter space.

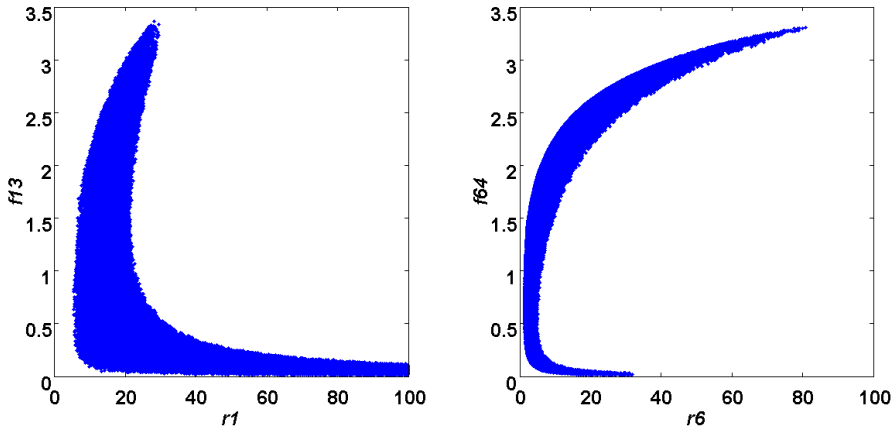


Figure C1. Two-dimensional projections of the viable parameter space onto the parameter axes of each independent flux (v_1: left, v_6: right).

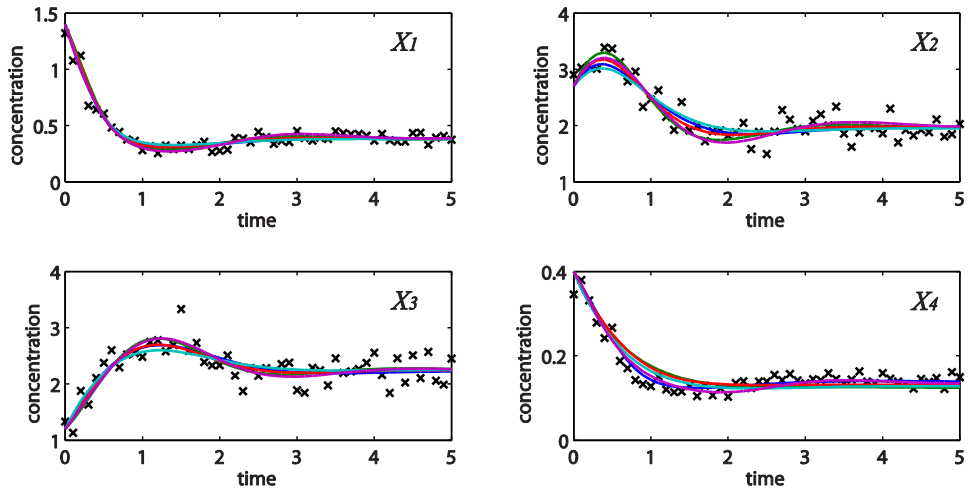


Figure C2. Concentration simulations of five randomly selected models from the ensemble (solid blue, brown, green, red and purple lines) versus the noisy data (×).

ACADEMIC PUBLICATIONS AND CONFERENCE PRESETATIONS

ACADEMIC PUBLICATIONS

– G. Jia, G. Stephanopoulos and R. Gunawan. Parameter Estimation of Kinetic Models from Metabolic Profiles: Two-phase Dynamic Decoupling Method. *Bioinformatics, Vol. 27 no. 14 2011, pages 1964–1970.*

– G. Jia, G. Stephanopoulos and R. Gunawan. Estimating Kinetic Parameters of Metabolic Networks within Flux-defined Subspace. *In Proc. of 8-th International Workshop on Computational Systems Biology (WCSB 2011), pages 96-99. Jun. 6-8, 2011, Zurich, Switzerland.*

– G. Jia, G. Stephanopoulos and R. Gunawan. Incremental Parameter Estimation of Kinetic Metabolic Network Models. *BMC Systems Biology, Vol. 6 no. 142 2012*.

– G. Jia, G. Stephanopoulos and R. Gunawan. Construction of Kinetic Model Library of Metabolic Networks. In *Proc of 8-th IFAC Symposium on Advanced Control of Chemical Processes (ADCHEM 2012), pages 952-957. Jul. 11-13, 2012, Singapore.*

– G. Jia, G. Stephanopoulos and R. Gunawan. Ensemble Kinetic Modeling of Metabolic Networks from Dynamic Metabolic Profiles. *Metabolites, Vol. 2 no. 4 2012, pages 891–912. (an invited contribution for the special issue on "Metabolic Network Models”).*

CONFERENCE PRESETATIONS

– ADCHEM 2012: International Symposium on Advanced Control of Chemical Processes (Singapore, Jul.2012);

– 11th and 12th International Conference on Systems Biology (U.K., Oct.2010; Germany, Aug.2011);

– 8th International Workshop on Computational Systems Biology (Switzerland, Jun.2011);

– 12th International Congress on Molecular Systems Biology (Spain, May.2011);

– 5th International Symposium on Design, Operation and Control of Chemical Processes (Singapore, Jul.2010).