DYNAMIC FLUX ESTIMATION

A Novel Framework for Metabolic Pathway Analysis

A Dissertation
Presented to
The Academic Faculty

by

Gautam Goel

In Partial Fulfillment
of the Requirements for the Ph.D Degree
Bioengineering in the
Wallace H. Coulter Dept. of Biomedical Engineering

Georgia Institute of Technology
DECEMBER 2009

DYNAMIC FLUX ESTIMATION

A Novel Framework for Metabolic Pathway Analysis

Approved by:

Dr. Eberhard O. Voit, Advisor
Wallace H. Coulter Dept. of Biomedical Engineering
Georgia Institute of Technology

Dr. Melissa Kemp
Wallace H. Coulter Dept. of Biomedical Engineering
Georgia Institute of Technology

Dr. Robert Butera
School of Electrical and Computer Engineering
Georgia Institute of Technology

Dr. Rachel Chen
School of Chemical and Biomolecular Engineering
Georgia Institute of Technology

Dr. Ana Rute Neves
Institute of Chemical and Biological Technology
ITQB, Portugal

Date Approved: August 19, 2009

To Mom and Dad

ACKNOWLEDGEMENTS

I would like to acknowledge Dr. Eberhard O Voit, my advisor and mentor, for his constant support and invaluable guidance in my research. Dr. Voit would push me (in the right direction) every time I felt I had reached a dead end. He has provided me enormous space and support to indulge in and explore diverse ideas. I have benefited immensely from his experiences and teachings in the past three years and will continue to do so for a long time to come.

I cannot thank enough two other people, who have been my colleagues and very close friends: Siren Veflingstad and Luis da Fonseca. Siren's support in validating my ideas has been instrumental especially in the early days when I conceptualized DFE. Her critique, on all my whimsical ideas and naïve enquiries, are perhaps best matched by her patience with me. Luis, on the other hand, helped me mature the ideas behind DFE when applying it to real-world scenarios. His expertise in biochemistry, and willingness to work on computational techniques have helped me shape DFE into its current form.

I would also like to extend my gratitude to all my committee members: Drs. Robert Butera, Melissa Kemp, Rachel Chen and Ana Rute Neves, for their support in my research and professional life. Each of the committee members has been extremely enthusiastic about my research. Their expert inputs have helped me significantly in meeting my research goals.

Finally, I cannot thank enough several my family members, both here in the US and back home in India, for their constant unconditional support. My parents, to whom I dedicate this work, have always had more faith in me than I myself. My brother, Anshul,

has been a true believer in my capacity for research and remains a die hard supporter of my career in science. My cousin, Amit S. Dhadwal, has been an inspiration, and an astute mentor. In his altruistic love and pride for his kins, I have always found wisdom whenever I have looked up to him. Lastly, I want to acknowledge the loving and caring support of my beloved wife, Kanu, who is now a driving force in my relentless pursuit of success.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

SUMMARY

High-throughput time series data characterizing magnitudes of gene expression, levels of protein activity, and the accumulation of select metabolites in vivo are being generated with increased frequency. These time profiles contain valuable information about the structure, dynamics and underlying regulatory mechanisms that govern the behavior of cellular systems. However, extraction and integration of this information into fully functional, computational and explanatory models has been a daunting task. Three types of issues have prevented successful outcomes in this inverse task of system identification. The first type pertains to the algorithmic and computational difficulties encountered in parameter estimation, be it using a genetic algorithm, nonlinear regression, or any other technique. The second type of issues stems from implicit assumptions that are made about the system topology and/or the functional model representing the biological system. These include the choice of intermediate pathway steps to be accounted for in the model, decisions on the irreversibility of a step, and the inclusion of ill-characterized regulatory signals. The third type of issue arises from the fact that there is often no unique set of parameter values, which when fitted to a model, reproduces the observed dynamics under one or several different sets of experimental conditions. This latter issue raises intriguing questions about the validity of the parameter values and the model itself. The central focus of my research has been to design a workflow for parameter estimation and system identification from biological time series data that resolves the issues outlined above. In this thesis I present the theory and application of a novel framework, called *Dynamic Flux Estimation (DFE)*, for system identification from biological time-series data.

CHAPTER 1

INTRODUCTION

The field of computational systems biology stands to benefit immensely from modern techniques in molecular biology that are able to generate comprehensive sets of multi-scale time-series data. These data generation capabilities, which once—but not anymore—lagged the methods of analysis and interpretation, drive biological systems analysis in two ways. Firstly, the quality and scale of data present an opportunity to steer biological systems analysis away from a purely descriptive mode to one that is complemented with rigorous mathematical modeling. This implies that systems biology must develop tools for providing objective and rigorous rationale for biological systems designs and modes of operation [1]. Secondly, the speed of data-generating technologies necessitates that computational systems biologists be able to incorporate newly available information quickly into their models and analysis. When done so, modeling will significantly accelerate the development and analysis of experimentally verifiable hypotheses leading to rapid scientific progress.

Concept Map Modeling and enabling software tools

Concept Map Modeling [2] is proposed here as a framework that facilitates the translation of heterogeneous biological information, whether explicit or intuitive, into mechanistic mathematical models. This novel conceptual framework bridges the gap between semi-quantitative biological knowledge and the construction of detailed mathematical models. The framework serves as a means to successful outcomes from the collaboration between biologists, who focus primarily on specific biological details and mechanisms, and computational systems scientists, who attempt to integrate diverse dynamic data into functional models for generating experimentally verifiable hypotheses.

Together, they enable the rich field of computational systems biology which promises to uncover the design and operational principles of biological systems through interactive information discovery. The foremost goal of this approach is the quantitative formalization of a biological phenomenon into a coarse mathematical model that integrates what biologists perceive to be functional systems surrounding their hypotheses. Subsequent steps allow analyzing and refining the rudimentary parametric models and connecting them with other coarse or detailed concept map models. Such coarse concept models, even when formalized with alleged local dynamic behavior of system components, provide significant insights for the biological systems [3]. However, the step to derive such models based on high quality time-series data is still beset with long computation times and questionable outcomes. In this thesis I present a software tool and computational techniques necessary to supplement the *Concept Map Modeling* framework. Collectively, these methodologies will facilitate rapid system identification and analysis based on biological time-series data.

The speed with which this framework can be effectively deployed however depends on two key enablers: *software tools* and *computational techniques*. For biological systems modeling and analysis to become a standard research technique with wider appeal, it is necessary not only to develop its theoretical foundation, but also to support all major methodologies with readily available, easy-to-use computational tools. We are still far from having such tools in a quality and accessibility comparable to modern word processors or spreadsheet programs, but the number of software packages for specific types of biological systems analyses is rapidly growing. New computational techniques become appealing to a wider audience only if they are supported by user-friendly software with an intuitive graphical user interface (GUI). Of course, many mathematical packages contain algorithms for integrating differential equations and for various types of optimization. Specific software has even been developed for solving and analyzing metabolic models, once they have been formulated in the form of fully

parameterized differential equations. Examples include PLAS[4], Gepasi[5], and BSTLab[6]. The packages BestKit[7] and Cadlive[8] furthermore allow the translation of topological diagrams into symbolic equations. With the motivation to realize a software tool, that would specifically support the activities and workflow of biological systems analysis within the frameworks of *Biochemical Systems Theory* and, in particular, of *Concept Map Modeling,* I developed the *Biochemical Systems Toolbox (BSTBox)* as the ***first specific aim of my research***. The framework of *Concept Map Modeling* and the associated software (BSTBox) are presented in chapter 2 of this thesis.

Biochemical Systems Theory (BST)

Biochemical Systems Theory (BST; [9, 10]) was originally designed for studying the dynamics and other features of biochemical and gene regulatory systems, but is not restricted to these application areas in terms of its mathematical foundation. BST is forty years old and its development, expansions, and applications have been documented in several books [11-14] and hundreds of journal articles, proceedings, and book chapters.

The basic tenets of BST are quite simple and translucent. In a nutshell, each variable that changes over time is given a name, typically X with an index, and its dynamics is formulated as an ordinary differential equation, which describes the change over time as it is governed by processes that affect this variable. The processes are functions of other variables within and outside the considered system and often of the variable itself. In most realistic cases, the modeler has some general information about these processes, but does not know their numerical details or even their mathematical structure. As an example, a population may grow in some sigmoid fashion, but the underlying mathematical function may not necessarily be known. BST addresses this problem by symbolically approximating each process with a product of power-law functions. One may wonder how it is possible to approximate something unknown, but this is conceptually no different than executing a linear regression on data points from a

system with unknown structure. The only difference is that the use of power-law functions in BST is much more general than linear regression. In fact, one can—quite surprisingly—show with mathematical means that any relevant nonlinearity can be faithfully represented in the power-law formulation of BST [15-20], which implies that we are not likely to run out of mathematical representations. On the biological side, many successful analyses attest to the validity of this approach (see, e.g., [12]).

To be specific, suppose that the variables of the system are called X_i and the processes V_i. In BST, each process V_i involving at most n dependent (state) variables and m independent variables takes the format

$$V_i = \gamma_i \prod_{j=1}^{n+m} X_j^{f_{ij}} \quad \text{..... Eq. 1}$$

Here, the dependent variables are governed by the dynamics of the system and change accordingly, whereas the independent variables are usually constant during each experiment and may include inputs, control variables and constant enzyme activities. Typical examples in a metabolic setting are metabolites (dependent variables) and enzymes or a (constant) substrate feed (independent variables). The rate constant γ_i describes the turnover rate of the process, and each exponent f_{ij} is a kinetic order that quantifies the direct effect of variable X_j on V_i. A positive kinetic order indicates an activating or otherwise augmenting effect, while a negative kinetic order reflects inhibition. The magnitude of each kinetic order reflects the strength of the effect. If there is no direct effect, the corresponding kinetic order is 0, and the variable, raised to 0, automatically drops out of the term, because any positive value of X, raised to 0, equals 1.

Equations within BST may be formulated in slightly different ways. In the Generalized Mass Action (GMA) representation, every process is considered individually. Thus, if the change in X_i is governed by p input and q output processes, the starting point is the equation

$$\dot{X}_i = [v_{1i} + v_{2i} + \ldots + v_{pi}] - [v_{i1} + v_{i2} + \ldots v_{iq}]$$

and each process v_{jk} is represented by a distinct product of power-law functions as shown above, so that the resulting equations always have the form

$$\dot{X}_i = \sum_{p=1}^{P_i} \left(\pm \gamma_{ip} \prod_{j=1}^{n} X_j^{f_{ipj}} \right), \qquad i = 1, \ldots, n \quad \textbf{.....Eq. 2}$$

As an alternative, the S-system form first aggregates all influxes for a given variable with a single power-law term and similarly aggregates all effluxes in a second power-law term. The starting point for this aggregation may thus be formulated generically for each variable as

$$\dot{X}_i = [v_{1i} + v_{2i} + \ldots + v_{pi}] - [v_{i1} + v_{i2} + \ldots v_{iq}] = V_i^+ - V_i^-$$

Again, the first group of terms in brackets consists of fluxes entering the metabolite pool X_i, the second group in brackets consists of fluxes leaving this pool, and V_i^+ and V_i^- are these groups of "aggregated" fluxes, which are now approximated as before. The corresponding S-system equations in BST therefore consist of at most one positive and one negative power-law term and have the format

$$\dot{X}_i = \alpha_i \prod_{j=1}^{n} X_j^{g_{ij}} - \beta_i \prod_{j=1}^{n} X_j^{h_{ij}}, \qquad i = 1, \ldots, n. \quad \textbf{.....Eq. 3}$$

Again, the rate constants α and β are non-negative and the kinetic orders g and h are real-valued. Variables that directly contribute to the influx into X_i are included in the

"alpha-term" and variables affecting the efflux from X_i populate the "beta-term". Variables that do not directly affect these terms have kinetic orders of 0 and therefore do not appear explicitly. The literature contains numerous descriptions of how to set up and analyze these types of models, as well as discussions about the similarities, differences, advantages and disadvantages of the two alternative representations (e.g., [11, 12]).

A key advantage of any representation within BST is that it is straightforward to set up equations from a diagram that shows how the components of the system affect each other. In fact, that is one of the key features of BSTBox, where given a list of components and their influences, the set of ordinary differential equations, describing the system dynamics, are automatically generated at the click of a button. Again, the challenging part is the identification of suitable parameter values.

Challenges in system identification from biological time-series data

Multi-scale time series data that are capable of characterizing temporal changes in magnitudes of gene expression, levels of protein activity, and accumulation of select metabolites *in vivo* are being generated with increased frequency and quality. These time profiles contain valuable information about the structure, dynamics and underlying regulatory mechanisms that govern the behavior of cellular systems. However, extraction and integration of this information into fully functional, computational and explanatory models has been a daunting task [2]. Three types of issues have prevented successful outcomes in this inverse task of system identification. The first type pertains to the algorithmic and computational difficulties encountered in parameter estimation, be it using a genetic algorithm, nonlinear regression, or any other technique. The second type of issues stems from implicit assumptions that are made about the system topology and/or the functional model representing the biological system. These include the choice of intermediate pathway steps to be accounted for in the model, decisions on the irreversibility of a step, and the inclusion of ill-characterized regulatory signals. The third

type of issue arises from the fact that there is often no unique set of parameter values, which when fitted to a model, reproduces the observed dynamics under one or several different sets of experimental conditions. This latter issue raises serious questions about the validity of the parameter values and the model itself.

Given a set of *in vivo* time-series data and a mathematical model in the form of a system of ordinary differential equations (ODE) describing its dynamics, the estimation of parameter values constitutes an inverse problem that many groups have attempted to solve using standard tools of optimization, system identification and intelligent searches. Notable methods include nonlinear regression [21, 22], genetic algorithms [23], evolutionary programming [24], and simulated annealing [25]. These methods have been the subject of extensive research in the past half-century, but till date there is no single method that performs well for all types of estimation tasks in biology. Optimization problems, in general, are concerned with locating maxima or minima of functions. When applied to the problem of parameter estimation, one usually begins with a hypothesized functional model and experimental data. The task at hand then is to estimate parameter values such that the model closely matches the experimental data when simulated. This involves constructing an *objective function* such as sum-of-squared-errors, beginning with an initial parameter guess, and then applying the optimization scheme of choice, which aims to minimize the error or difference between the data generated by the simulated model and the observed experimental data. It is now widely acknowledged that this seemingly straightforward approach is in fact riddled with many complications. Firstly, it is an iterative process requiring very many evaluations of an ODE model that usually consumes more than 95% of the computation time and can be extremely slow irrespective of the optimization algorithm [26]. Secondly, the numerical solution to the ODEs is frequently confounded by a selection of parameter values that may cause the solver to estimate negative or unreasonably high values for concentrations. Lastly, the performance of the local optimizer is greatly impeded by the dimensionality of the

problem, the constraints imposed along each of the dimensions, restrictions on the error tolerance during the search and choice of the number of iterations for which the optimization is performed.

My previous research [27] supported the often-mentioned notion that ODE-based local optimization guarantees success when initiated with a reasonably "good" parameter guess, one that is inside the *basin of attraction* of a local or global minimum. Over the years, I have explored numerous estimation schemes that had the potential to lead to such "good" initial guesses while providing guidance on the *error landscape* given a model structure, time series data, an objective function and choice of an optimization scheme. These schemes included artificial decoupling of equations without precursor-product constraints, multiple shooting for system integration, and slope-based parameter estimation, to name a few. These different schemes worked with varying degrees of success and the problem of parameter estimation could be solved when I applied these schemes collectively in different combinations. Success, however, came only after months of heuristic search. The general problem of parameter estimation was clearly not solved.

In other (unpublished) work I tried to approximate the *manifold of good solutions* in parameter space and its *basins of attraction*. I considered two sets of parameter values Ω_A and Ω_B. These sets could either fit a model to the same experimental data set 'A' or to different data sets 'A' and 'B' simultaneously. The hypothesis was that in case of the former, the solutions (parameter sets) would lie along a continuum (*manifold*). In the latter case, a common solution to the two experimental data sets would lie at the intersection of two such manifolds. Further research revealed that it was extremely difficult to make any reliable conclusions about the parameter space, which was presumably due to the *curse of dimensionality.* Any attempts to approximate the functional form for a hypothetical manifold were confounded by the scale of the problem at hand. Not to mention, this in turn, became more or less like any other system

identification problem. A reason for the lack of success might be the observation of Greenside *et al.* [28] who showed that even the simplest Newton line-search algorithm, when applied to solving a moderately complex problem (in their case, the polynomial $z^3 - 1 = 0$ in the complex plane), exhibits an intricate fractal basin of attraction of the roots. Kutalik *et al.* [29] have demonstrated the presence of such *basins of attraction* in higher dimensions whereby the manifold was approximated in two-dimensions at a time. My efforts to approximate the manifold as a *hyperplane* in higher dimensions have borne no success to date.

Aside from these computational complexities, the second type of issues stems from the assumptions made about the underlying system topology and/or the representative functional model. Voit *et al.* [30] have discussed such issues at length. They include the choice of intermediate pathway steps to be accounted for in the model, decisions on the irreversibility of a step, and the inclusion of ill-characterized regulatory signals. Oftentimes, the assumed kinetic functions themselves are questionable and there have been no means to either establish the appropriateness of a specific functional form of choice or to evaluate its efficacy with alternative models. This poses a serious threat to the "goodness" of a model because the assumptions pertaining to the underlying functional flux severely limit the reliability of its predictions in untested experimental conditions. Moreover, if two different models are able to explain the same experimental data, then there are no criteria to objectively compare or distinguish these two models.

Lastly, I have also found that independent of the choices of parameter estimation schemes, the system topology, the underlying functional model, and even the modeling mode (bottom-up/top-down/partial) there is a more fundamental problem of a different nature, to model "fitting". This third type of issue arises from the fact that there is often no unique set of parameter values, which when fitted to a model, reproduces the observed dynamics under one or several different sets of experimental conditions. At first this may rather seem to be an advantage for parameter estimation since it widens the spectrum of

optimal solutions. And that in fact is true in specific situations, *e.g.*, when estimating parameters for a reversible step in a pathway. It can be easily shown that different numerical combinations of parameter values will be able to reproduce the local dynamics of such a step. However, when more than one set of parameter values are able to fit an entire pathway model and reproduce the observed dynamics then it lowers the confidence in these parameter values. Moreover, when a single set of parameter values is unable to yield the appropriate dynamics in different experimental conditions then it raises doubts, not only on the parameter values, but also on the underlying functional model, the system topology and possibly even the data themselves. In such situations, conventional wisdom suggests that one should try to search for an optimal set of parameters that will fit the model to *all* experimental data sets simultaneously. To date, however, there is no guidance or diagnostics available to resolve the dilemma of whether the true problem lies in the model, the topology, the data, or combinations of the three.

With the motivation to resolve these issues, I set as ***my second specific aim of research*** the task to design a novel approach to parameter estimation and system identification. To meet my objectives I have researched the application of a systems analysis technique that has become quite popular, namely, "decoupling with slope estimation" [31]. Several authors have suggested different variations of decoupling systems of differential equations that suited their particular purposes best. Voit and Savageau suggested decoupling by estimating slopes of all (n) time courses at many (N) time points, which reduced the system of n coupled differential equations to n X N algebraic equations [13, 26, 31]. Voit *et al.* [32] have applied decoupling as a means to validate implicit or explicit model assumptions. Chou, Martens and Voit [33] have successfully designed a parameter estimation scheme for S-systems (*Alternating Regression*) based on the same principles.

When trying to apply *Alternating Regression* for parameter estimation with GMA-models, I encountered a major challenge, which may be called "*incongruent flux*

compensation", that I believe would be a characteristic of any decoupling-based estimation scheme. I briefly discuss the issue here. In any metabolic pathway, the efflux from one metabolite pool (precursor; $\boldsymbol{X_i}$) should equal the influx into the next metabolite pool (product; $\boldsymbol{X_j}$), assuming there are no secondary pathways that may divert material from the primary pathway and/or add material to it [30]. These relationships are known as "precursor-product constraints" and it is imperative that any model, even when decoupled, should meet these constraints to maintain mass balance. This, however, poses a unique challenge to the parameter estimation scheme that, as per the decoupling approach, attempts to solve the system one equation at a time. When estimating for parameters for the equation describing the dynamics of the precursor (X_i), parameters are fitted to its production and degradation terms simultaneously. In such a scenario, these parameters are optimized only to yield the necessary flux compensation between the production and degradation functions of X_i such that these accurately capture the net dynamics of X_i. However, the efflux of precursor (X_i) so fitted may be suboptimal as an influx to the product (X_j). I have argued that this fundamental issue of compensation between uncertain fluxes can be circumvented by first deriving time-series profiles of all fluxes within the system, intra- and extra-cellular, and then estimating for parameters one flux at a time. I have designed a scheme for the successful application of slope-based decoupling method to obtain such "model-free dynamic flux estimates". A complete overview of this novel approach to parameter estimation, called *Dynamic Flux Estimation (DFE)*, and preliminary results from its application to a proof-of-concept metabolic pathway are presented in chapter 3 of this thesis.

Significance of model-free dynamic flux estimates

A priori knowledge of dynamic intra- and extra- cellular flux profiles, derived from time series data, can have far-reaching implications for parameter estimation and system identification at large. Several groups have realized the potential of NMR and

mass spectroscopy to examine intracellular fluxes and are combining experimental and analytical methods to study flux distributions in metabolic networks [34-38]. At present, these analyses are always constrained by steady-state or pseudo-steady-state assumptions. The proposed *Dynamic Flux Estimation (DFE)* approach does not make any such assumptions. Furthermore, when applied correctly, this approach yields time-series profiles of all intra- and extra-cellular fluxes in a system without any assumptions regarding the underlying functional model.

There appear to be numerous advantages with the successful application of this approach. These can be categorized into benefits for parameter estimation, model identification and model evaluation. When applied to parameter estimation, the use of the decoupling technique based on slope estimation keeps this approach computationally inexpensive. Given the time-series data for all substrates, products and co-factors and the flux time-series for each reaction step in a pathway, one is able to undertake a rigorous investigation of the most appropriate functional form that can accurately capture the observed dynamics. The evaluation of flux-substrate plots is a first step in this analysis. Furthermore, well researched tools of statistical analysis, such as jack-knifing, bootstrapping etc. can be used to further assess the suitability of each alternative functional form. Thus, a truly objective and rational assessment renders it possible to choose between representative kinetic forms (Mass-action / Michaelis-Menten / Power-law / Hill equation etc.). Furthermore, inherent complexities of parameter estimation will be greatly reduced. For instance, if using the power-law formalism, the ultimate estimation of parameters for each flux translates into a simple linear regression problem in log space. Finally, this approach to parameter estimation would overcome the earlier stated fundamental challenge in decoupling-based estimation schemes: "*incongruent flux compensation*", which never arises with the proposed approach.

Besides ameliorating the parameter estimation process, the dynamic flux-based approach addresses many issues in model identification. Foremost it clearly separates the

sources of error that reside in inconsistent data from those that arise due to invalid assumptions about the system topology and specific functional forms. These issues are circumvented because, as a first step, this approach enforces consistency checks among the experimental data, system stoichiometry and the hypothesized topology. Subsequently, the dynamic intra- and extra-cellular fluxes in the system are derived without any assumptions regarding the underlying model, i.e. the functional kinetic forms. All assumptions of the system topology, such as decisions on irreversibility or even the inclusion or exclusion of a step, are validated at this stage itself. Next, as explained earlier, it is possible to objectively assess the appropriateness of alternative functional forms. Lastly, when fitting the parametric kinetic forms to these fluxes, the assumptions of ill-characterized regulatory signals and co-factors will be validated by the ability of the model to reproduce the derived flux dynamics. Moreover, the "goodness" of a fully parametric model, obtained from such a staged modeling process, is directly evident in the outcomes of each stage – the mass and flux based balances, the biological plausibility of the derived flux profiles, the fit of the functional forms and the corresponding parameter values.

Ultimately, one will be able to rationalize the predictive power of a model (or the lack of it) and consequently establish its "value", which is typically assessed from its ability to reliably extrapolate toward untested conditions, by independently evaluating the data, the topology and the representative functional form. Before using a fully parametric kinetic model to predict situations that had not been used for model identification or data fitting, one will be able to test whether the particular, chosen parametric model is suitable for those test conditions or not. Different functional forms may be justified in fitting the same data as long as they are consistent with the underlying flux profiles. Likewise, different parameter sets that fit the same data may also be acceptable as long as they reproduce the same flux dynamics. In the event that model predictions do not match the experimental test data, one will be able to assess and justify minimal changes that need to

be made to a model (parameters or functional form or system topology) by analyzing the new data with the same dynamic-flux based approach.

To substantiate the above stated claims, I extended the DFE approach from its preliminary proof-of-concept form to a set of working methods that would help gain a wide-appeal and application. As ***my third specific aim of research***, I have demonstrated DFE to be a robust approach when applied to real-life situations of noisy, incomplete time-series data with under-determined linear system of fluxes. These results are presented in chapters 3 & 4 of this thesis.

Real-life scenario: Regulation of glucose metabolism in *L. lactis*

Lactic acid bacteria have a long tradition in industrial fermentations, where they are used as starters in the manufacture of fermented foods and beverages, such as buttermilk, cheese, and yogurt, sausages, bread, pickles and olives, and wine. In particular, *Lactococcus lactis* is widely used in the dairy industry for the production of cheese and buttermilk, mainly due to its capacity to convert about 95% of the milk sugar lactose to lactic acid. The low pH generated by this activity inhibits the spoilage and growth of pathogenic bacteria, and consequently extends the shelf-life of the fermented products. The relative simplicity of *L. lactis* metabolism that converts sugars via the Embden-Meyerhof-Parnas pathway to pyruvate and generates energy mainly through substrate level phosphorylation, together with a small genome with limited redundancy, and a multitude of genetic tools, make this organism a very attractive model for systems biological approaches. Regulation of glycolysis in *L. lactis* has been the subject of intensive research over the past three decades. Key enzymes in the homofermentative pathway, phosphofructokinase, fructose 1,6-bisphosphate aldolase, glyceraldehyde 3-phosphate dehydrogenase (GAPDH), pyruvate kinase (PK) and lactate dehydrogenase (LDH) were characterized, and concentrations of several glycolytic intermediates in cell extracts had been obtained already in the eighties. However, despite the wealth of

metabolic information collected, a comprehensive understanding of sugar metabolism and regulatory pathways in this model organism has not been achieved yet. (For a complete review see [39]). ***As my fourth specific aim of research*** I used the DFE approach to analyze the metabolism of wild-type non-growing cells of *L. lactis* under anaerobic conditions. Using an integrative top-down and bottom-up approach to system identification, facilitated by DFE, I deduced the dynamic flux profiles which when evaluated for multiple data sets revealed distinct patterns of regulation. Using these dynamic flux data, I developed a detailed kinetic model by combining time-series data (of metabolites, cofactors and fluxes) with kinetic and regulation information obtained from independent *in vitro* enzymatic studies. The details of model development and analysis are presented in chapter 5 of this thesis.

CHAPTER 2

CONCEPT MAP MODELING [1]

It is proposed that computational systems biology should be considered a biomolecular technique of the 21st Century, because it complements experimental biology and bioinformatics in unique ways that will eventually lead to insights and a depth of understanding not achievable without systems approaches. This chapter begins with a summary of traditional and novel modeling techniques. In the second part, it proposes *Concept Map Modeling* as a useful link between experimental biology and biological systems modeling and analysis. *Concept Map Modeling* requires the collaboration between biologist and modeler. The biologist designs a regulated connectivity diagram of processes comprising a biological system and also provides semi-quantitative information on stimuli and measured or expected responses of the system. The modeler converts this information through methods of forward and inverse modeling into a mathematical construct that can be used for simulations and to generate and test new hypotheses. The biologist and the modeler collaboratively interpret the results and devise improved concept maps. The last section of the chapter describes software, *BSTBox*, supporting the various modeling activities.

Biological Systems Analysis

While bioinformatics serves an extremely valuable purpose, it is by itself not sufficient a computational tool to yield true understanding of how biological systems

---

[1] Part of this chapter is published in: G. GOEL, I.-C. CHOU and E. O. VOIT, "Biological systems modeling and analysis: A biomolecular technique of the twenty-first century". J. Biomol. Tech. 17(4): p. 252-269, 2006

function. Experimentation has given us a more complete parts list than was ever available before, and bioinformatics has allowed us to sort and manage this list with admirable efficiency. However, what is still missing is a set of tools that explain the rationale of a given biological component; that can determine with objective means why a particular, observed design in nature is superior to other designs that at first appear to be just as reasonable; that can merge diverse data and contextual pieces of information into quantitative, conceptual structures that can be analyzed with the rigors of the universal language of mathematics. The field of biological systems modeling and analysis, or "computational systems biology", addresses these tasks.

Biological systems analysis can be traced back to several roots. Some of these may even be seen in the holistic views of antiquity [40] or found in the early work of physiologists (*e.g.*, see [41]), who long ago began to investigate the nervous *system*, the digestive *system*, the cardiovascular *system* and many other complex structures in the human body as integrated entities, in which diverse components had specific roles yet worked together synergistically to achieve much greater tasks than what each component could have accomplished on its own. It has also been suggested to place the origins of systems biology by considering it as the evolution of molecular biology into genome-wide analyses (*cf.* [42]).

While such assessments have legitimacy, more refined definitions of biological systems analysis and modeling require that purely descriptive approaches to biology be accompanied by the ability to make reliable, quantitative predictions of the responses of cells or organisms to experimentally untested situations. Moreover, as stated before, biological systems analysis must develop tools for providing objective and rigorous rationale for biological system designs and modes of operation [1]. The need to make predictions and to discover general design and operating principles necessitates the consideration of a different set of roots from which modern biological systems analysis draws. This additional heritage is theoretical in nature and embodies ideas and concepts

that reach beyond the possibilities afforded by traditional biological experimentation. As was true for the glorious era of physics in the early 20th century, we are beginning to recognize the necessity to support biology with a rigorous mathematical foundation. Such a foundation not only permits bookkeeping of the large, functional assemblages of heterogeneous molecules that we encounter everywhere in biology, but is a prerequisite for the formulation of rules and general laws that will eventually form the rudimentary building blocks of a theory of biology. The roots of this crucial aspect of systems biology are evident in the work of seminal thinkers and visionaries like von Bertalanffy, who over half a century ago proposed the mathematical characterization of organisms as dynamic, open, nonlinear, complex systems [43]. The work of von Bertalanffy, Lotka, Wiener, Weaver, Turing, Rosen, and many others who strove toward the quantitative description of biological systems and the discovery of general principles governing biology shaped the theoretical foundation of today's computational systems biology and will continue to provide guide posts for its future.

Biological systems analysis will never replace hypothesis-driven and reductionistic biological research, but constitutes its conceptual and practical complement. Biological systems analysis may ultimately reach a role as central as it is in physics, where experiments are only executed after their theoretical underpinnings have been proven beyond doubt, but we are presently far away from such prominence. In the meantime, biological systems analysis will provide additional tools and techniques for functionally organizing diverse pieces of information and data that stem from traditional biological experimentation. Therefore, computational systems biologists must collaborate closely with experimentalists who focus sharply on select biological details and mechanisms.

Biological systems analysis may be dissected into four components:

1. Development and application of high-throughput methods of biological data generation and quantitative analysis with the objectives of detecting and

identifying unknown biological system components and of creating a comprehensive catalog of these components, along with their roles and interactions;

2. Integration of biological information from *de novo* experiments and the literature into functional contexts through the creation of conceptual, mathematical, computational, and informational models that relate the multitude of molecular components to each other both within and among different levels in the hierarchy of biological organization, such as the genomic, proteomic, metabolic, and physiological levels;
3. Specific experimental testing of hypotheses generated through mathematical and computational modeling; and
4. Approaches toward a true understanding of the design and operation principles of small and large systems in biology through information discovery and through the identification of the specific systemic roles that the components of these systems play, their connectivity, their influences on each other, and their synergisms.

The remainder of the chapter describes three aspects of computational systems biology. First, some of the traditional methods and techniques of biological systems modeling and analysis are reclassified. In the following section, a variation on these methods is introduced, namely the novel technique of *Concept Map Modeling*. Finally, a preliminary software is demonstrated that supports traditional and concept map modeling.

Forward, Inverse and Partial Modeling

To the uninitiated, mathematical modeling is often seen as one standard set of tools, conceptually similar to a specific technique like electron microscopy or microarray analysis. Indeed, experimentalists frequently approach a modeler with the request to

"model their data." The truth is that mathematical modeling comprises an enormous repertoire of techniques, and the only real commonality is that they all lead to some mathematical representation of a biological phenomenon (the "model"), which is subsequently analyzed and interpreted in biological terms. To some degree, the type or classification of the mathematical representation is a technical issue. Thus, a model may be deterministic or stochastic, continuous or discrete, mechanistic-explanatory or more like a black box. Irrespective, the modeling *process* may be exactly the same. First, a symbolic model is constructed from first principles like physical laws, as an extension of an already existing model, or from intuition. This model almost always consists of equations that contain variables and parameters. Variables could be plant or animal species in an ecological system, metabolites in a pathway model, or the expression levels of particular genes in a genome experiment, while the parameters describe more or less fixed quantities like the reproductive rate of a species, the K_M of an enzyme, or the transcription rate between DNA and RNA. The analysis of the mathematical model usually requires knowledge of all parameter values, which therefore need to be identified from the body of biological knowledge. While this may sound like a straightforward task, the estimation of parameter values is often very challenging and has been, and will continue to be, the most daunting bottleneck of mathematical modeling in biology. Once the parameters are estimated, the analysis of the model is (one could say 'merely') a matter of mastery of the tricks of the trade of mathematics and computer science. Because of the complexity of biological systems, the analysis is typically executed per computer, sometimes with elegance, but more often with brute force, grinding out approximate solutions that are more than sufficiently accurate for most biological purposes. The interpretation of results is ideally performed in collaboration between the subject area biologist and the mathematical modeler or computer scientist. The real obstacle to fast progress in biomathematical modeling is thus the determination of

unknown parameters from biological information and three classes of methods that are available for this task are described here now.

Forward Modeling. The standard way of identifying parameters is based on "local" information. Thus, to construct a model of a metabolic pathway, one considers one enzymatic or transport step at a time, combs the literature for information about this enzyme, its cofactors and modulators, and translates this information into a mathematical rate law (such as the v_{jk} above), which could be a Michaelis-Menten or power-law function, among a wide variety of possibilities. The collection of all rate laws governs the dynamics of the model. Comparisons of the model responses with biological observations support the validity of the model or suggest adjustments in assumptions or parameter values. If done right, this forward process can ultimately lead to model representations of the pathway that exhibit the same features as reality, at least qualitatively, if not quantitatively. Some BST examples are models of the TCA cycle [19], purine metabolism [15-17], the citric acid cycle [20, 44], the Maillard-glyoxylase network with formation of advanced glycation end products [45], the trehalose cycle [46], sphingolipid metabolism [47, 48], the ferredoxin system [49], and the regulation of glycolysis [30, 50]. In almost all of these cases, the strategy consisted of setting up a symbolic model, estimating local parameters, studying the integration of all individual rate laws into a comprehensive model, testing the model and making refinements to some of the model structure and the parameter values.

The main disadvantage of this strategy is that a considerable amount of kinetic information is needed, but that this information is often only available from differently structured experiments and often only from different species. Furthermore, the process of construction and refinement is very labor-intensive and requires a combination of biological and computational expertise that is still rare.

Inverse Modeling. Modern molecular biology is offering an alternative in the form of high-throughput experimental data. Particularly useful for modeling and

parameter estimation are measurements of biological components (metabolites, proteins, gene expression) at a series of time points after a stimulus. As an example, which will be use throughout for illustration, Neves *et al.* [34, 51, 52] used *in vivo* NMR techniques to measure the concentrations of most glycolytic metabolites in the bacterium *Lactococcus lactis*. This type of data allows, at least in principle, an entirely different path toward suitable parameter values. Namely, the symbolic BST model (without specified parameter values) is fitted to the time series data by means of an optimization algorithm. Thus, in contrast to the forward, "bottom-up" approach described before, parameters are estimated from the observed data "globally" or "top-down." The advantages are manifold. Most important is that the data come from the same organism, are obtained under the same well-characterized experimental conditions, sometimes even *in vivo*, and therefore account for all processes within the organisms that could have an effect on the variables of the system (*cf.* [32]). A significant disadvantage of this strategy is that the estimation process itself is very challenging computationally. Also, many of the processes that affect the dynamics of the system *in vivo* are often either unknown in detail or not even considered at all in the model. As a typical example, most standard models of glycolysis show glucose 6-phosphate as a simple intermediate in a linear chain between glucose, fructose 6-phosphate and fructose 1,6-bisphosphate. In reality though, glucose 6-phosphate is in equilibrium with glucose 1-phosphate and also serves as the initial substrate for the pentose shunt.

The inverse strategy may also be used for time series data that stem from a pathway with a structure that is not fully known or whose regulation is obscure. At least in principle it is possible to estimate parameter values from the data and interpret them as structural and regulatory features. With good data, this is presently possible for relatively small pathways [26, 33, 53].

Partial Modeling. A particular problem with any modeling approach arises in the form of "ubiquitous" metabolites like ATP and NAD. Again using the example of

glycolysis, ATP and NAD are clearly important players, but they are also involved in dozens of other reactions that are not part of the model. In the past, some mathematical models have considered them constant, which really defeats the purpose of glycolysis. Others have employed conservation relationships between ATP, ADP and AMP or NAD and NADH, thus allowing for some dynamics without having to model a lot of details (*e.g.*, [54]). Another alternative, gleaned from biochemical lab experimentation, is the construction of mathematical buffers that absorb excess material, while providing material in times of high demand. The mathematical features of the buffers can be designed to adjust for dynamic variations in concentration at a predetermined rate [55]. As yet a different manner for dealing with ubiquitous metabolites, a partial modeling approach has been proposed in the past [56], which allows one to mix well-defined components with components whose dynamics is only known in the form of time series that are experimentally observed but cannot be formulated in terms of other model components. As a case in point, for the analysis of time series data describing pyruvate and lactate production in *L. lactis* [30] measured ATP, P_i, NAD and NADH concentrations over time were available, but it is extremely hard to formulate their dynamics as functions of the system variables, because each of these factors is involved in dozens of reactions, most of which were not modeled. As a solution, the better-defined components were formulated as differential equations in BST, and their dynamics included ATP and NAD as variables. However, ATP and NAD were not modeled as a differential equation *per se* but entered the system as time-dependent input functions. In mathematical jargon, the observed data were considered as raw or smoothed "forcing functions" to the differential equation system.

Development of Concept Map Models

While the traditional and newer modeling strategies are very valuable, an important step is missing between the wet experiment, biological insight and intuition, and the construction of mathematical models. This step consists of the translation of "heterogeneous" biological diagrams into symbolic and subsequently parameterized equations. "Heterogeneous" means that some components in such diagrams refer to metabolites, some to genes, and some to a variety of processes, such as apoptosis, differentiation, or the initiation of a disease process. A good example is illustrated in Figure 1, which shows how surface features can trigger proliferation and differentiation events in overlaying cells. "Concept map modeling" attempts to fill the gap between this type of mixed explicit and implicit information and the typical mathematical models encountered today in biology.

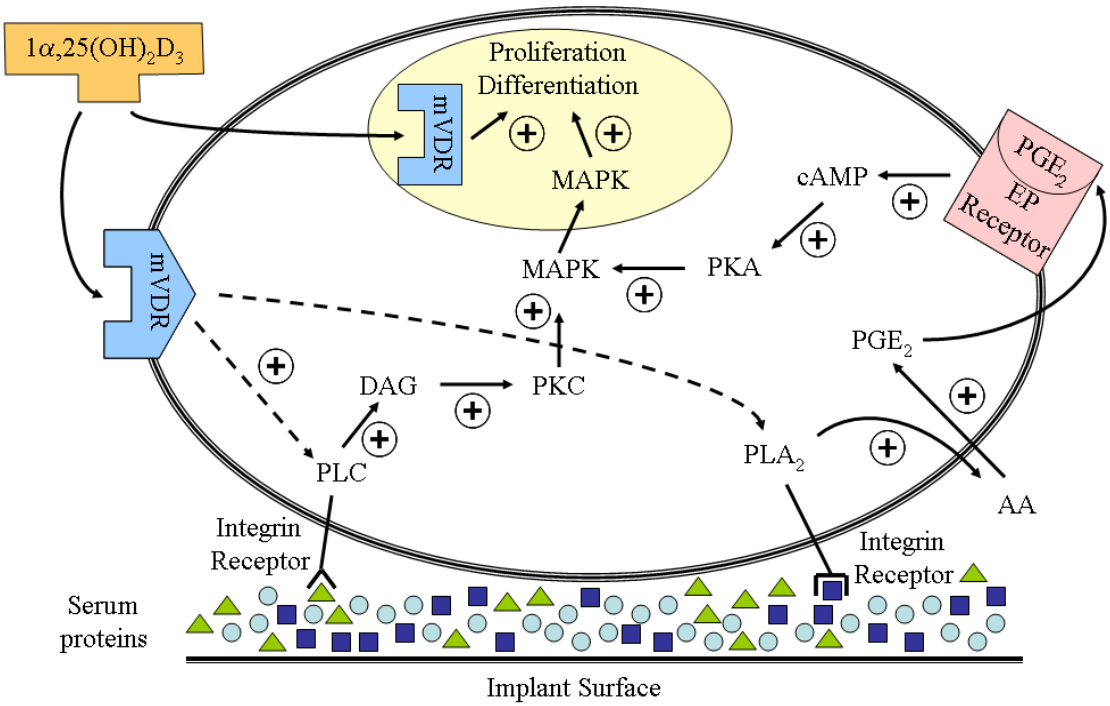


Figure 1: Concept map summarizing siRNA knockdown studies demonstrating how effects of the physico-mechanical environment are mediated by integrins. (Barbara Boyan, pers. comm)

The goal of concept map modeling is the development of novel methods for formalizing qualitative knowledge in diagrams like Figure 1, for creating conceptual models, and for making these amenable to coarse, and later refined, mathematical analysis. Typical conceptual models involve components and processes at various levels or biological organization, and the traditional response to this situation by the modeling community has often been to set up models within each level, with the implicit or explicit purpose of connecting them at a later time (e.g., [41]). To reach its ultimate goals, this approach requires a lot of time, and it does not even fully exploit the implicit knowledge that biologists associate with such concept maps. It seems therefore beneficial to derive models that directly capture the biologist's intuition and permit the inclusion of incomplete and heterogeneous information from one or several levels at once. Thus, the foremost goal of this approach is it "to get started" with a quantitative formalization of a biological phenomenon, by integrating in a coarse mathematical model what biologists perceive to be functional systems surrounding their hypotheses.

The approach begins with the translation of a hypothesized, static biological map into a hypothetical, mathematical model structure. This step is followed by the formalization of observed or alleged local dynamic responses of system components under a defined set of external conditions. These two ingredients allow one, at least in principle, to infer a fully parameterized model, which in subsequent steps is analyzed, refined, and connected with other coarse or detailed concept map models. A flow diagram of this approach is shown in Figure 2.

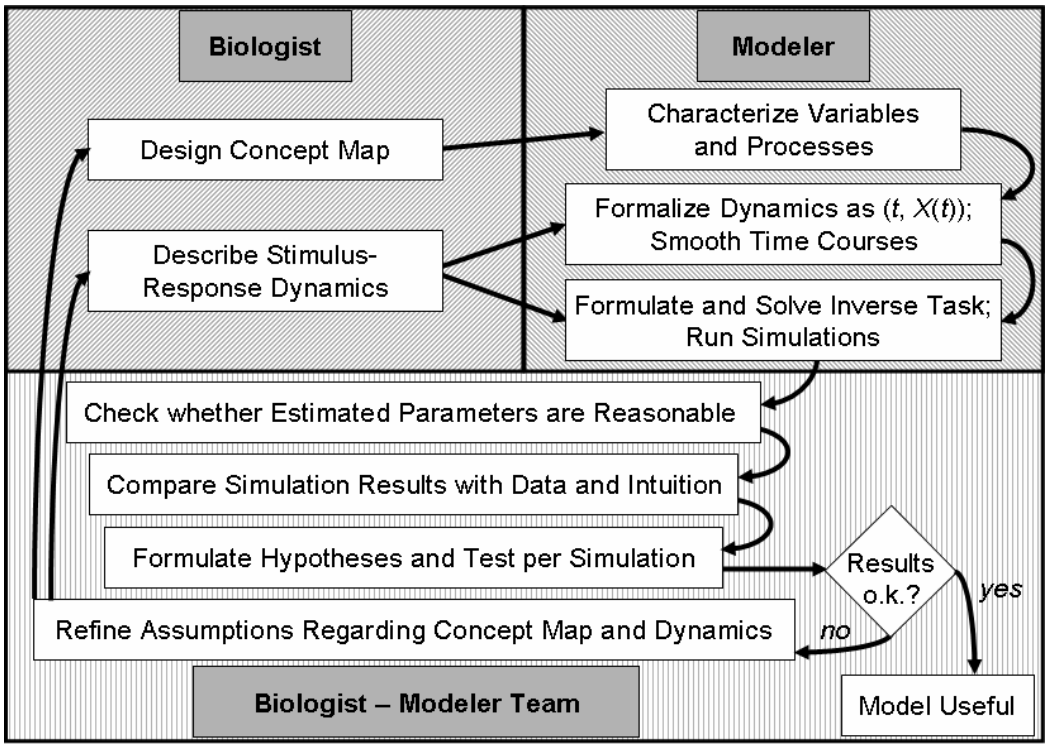


Figure 2: Flow diagram of the proposed approach to formalizing biological concept maps

The initial step of this effort is to establish, in collaboration with biologists, lists of components and processes with relationships and rules that are visualized in the concept map. At this stage, it is necessary to discuss, question, and revisit in detail how the pieces within each aspect of the diagram conceptually relate to each other and how they contribute to the overarching functional entity, which at the highest level will eventually become a comprehensive model of the topic of investigation. In very many cases, these maps already exist in the heads of experienced biologists, and sometimes they had already been sketched out and published. However, the maps themselves do not allow further, quantitative analysis, and it is therefore desirable to facilitate the translation of hand waving arguments into mathematically testable structures and analyses.

For each component of the map, one records details that might become important, as well as the biologist's level of confidence in all pieces of qualitative and quantitative

information populating the map. To mathematically trained scientists, this first formalization step might be uncomfortable and appear almost unscientific, but it is a crucial (and ultimately very rewarding) process, because biological systems analysis cannot wait until all details of a system are known and solidly quantified. To ensure that the mathematical representations of concept maps are consistent and unambiguous, it is useful to develop consistent diagramming conventions for concept maps that extend upon ideas for biochemical maps, as well as a controlled vocabulary used to capture the map and its associated semi-quantitative data. Choosing BST as modeling framework, the structural and regulatory information from traditional maps is already sufficient to set up symbolic equations. Indeed, this step is routine textbook material and can be accomplished with computer software (see below).

Biologists almost always know much more than what is captured in static maps. For instance, they often have at least some knowledge of the types of reactions that are possible in a single enzymatic step or the time it takes between a change in gene expression and the corresponding change in enzyme activity. This knowledge enables the biologist to convert the static concept map into a collection of Boolean or semi-quantitative dynamic (SQD) maps for given scenarios. In the Boolean case, a typical statement like "if we knock out gene X, then process Y does not occur" aids in the refinement of the static map, because it suggests the closer consideration of direct and indirect influences of components inside or outside the system. In the SQD case, a typical observation may be "if we bathe the cells in a medium containing A, B starts to rise within four of five minutes and C decreases to about half its normal level". In the ideal case, one would be able to determine from this information accurate functions for each node over a range of scenarios, which would permit direct application of inverse methods [30, 57], as discussed before.

In realistic cases of concept map modeling we will usually not have detailed time series but rather semi-quantitative or only qualitative information on the dynamics of

each node. Still, one can use this minimal information for the construction of a coarse initial model. Substituting for actual time series, one may choose a simple function that captures the dynamics at each node, according to the biologists' observation, general experience, or intuition. The particular mathematical choice of these functions is not all that critical, because they are only used for the inverse task in lieu of smooth data. Thus, the emerging software (see below) permits the biologist to sketch the process dynamics, for instance, in the form of a saturating or sigmoidal curve. Once drawn, the software permits modifications and alterations in this shape by allowing the addition of points or free dragging of the curve. These curves capture the biologists' semi-quantitative understanding of the effect of a given variable on the dynamics of a given node in a given scenario.

Once the alleged dynamics of all variables is entered, the curves are treated like measured time series and methods of inverse modeling can be used to estimate parameters. One must caution that, in contrast to the relatively straightforward symbolic model design step, this parameterization is complicated. While the current software does permit parameter estimation, it is still much too slow for an interactive exploration of the concept map, and further research will be necessary to make this step fast enough. Several methods have lately been proposed for this purpose, but none of them is at this point sufficiently reliable, stable and efficacious.

In summary, the starting point for concept map modeling is a network model based on the known or hypothesized connectivity and regulatory information on the static concept map (forward step), as well as a set of assumed or manually entered functions that mimic the observed or hypothesized dynamic responses under a specific scenario at each node of the network. These two ingredients are sufficient for inverse modeling techniques to estimate parameters of a global BST model for one scenario at a time or for a collection of several scenarios that are assessed simultaneously. The resulting coarse models may be used in two ways. First, it is possible to test stability, sensitivities, gains,

and other features that are routinely studied in biological systems analysis [11, 13]. These features show whether the conceptual model has a chance of being correct, because very high sensitivities, or lack of stability, are often unrealistic. Secondly, one can now run simulations, which are cheap to execute and often quickly lead to the discovery of weaknesses in the model or confirm assumptions made. The results of this set of analyses are shared with the subject area biologists for interpretation and revisions of their concept maps, thereby initiating the typical iterative cycle of modeling and validation (*cf.* Figure 2).

Software

New computational techniques become appealing to a wider audience (only) if they are supported by user-friendly software with an intuitive graphical user interface (GUI). Of course, many mathematical packages contain algorithms for integrating differential equations and for various types of optimization. Specific software has even been developed for solving and analyzing metabolic models, once they have been formulated in the form of fully parameterized differential equations. Examples include PLAS [4], Gepasi [5], and BSTLab [6]. The packages BestKit[7] and Cadlive [8] furthermore allow the translation of topological diagrams into symbolic equations. Shown here is an interactive MATLAB® module that is under development. This software, *Biochemical Systems Toolbox (BSTBox)*, is presently available as a preliminary test version that allows the user to conduct traditional and concept map modeling, as it was described above. The main features for this toolbox are:

1. interactive generation of lists of variables, along with their influences;
2. interactive preparation of measured or hypothetical time courses;
3. automated formulation of BST equations, in accordance with the lists in (1);
4. estimation of parameter values from time course data in (2);
5. generation and visualization of computed time courses;

6. testing of inferred and alternative models.

Besides the basic MATLAB installation, *BSTBox* also requires the Optimization, Genetic Algorithm and Spline Toolboxes; future versions may reduce these requirements. *BSTBox* is designed to facilitate forward, inverse, partial, and concept map modeling. Each major step of the modeling approach is supported by appropriate functionality designed in separate "tabs" in the toolbox, which unlock progressively as the user progresses through the stages of model development and analysis.

Upon successful installation of the required libraries in MATLAB, *BSTBox* is invoked with the simple command 'BST' which allows users either to: 1) create a new model (map) from a list of components and time course data (experimental or hypothetical) specified in a Microsoft Excel file; or 2) load an already existing model (map) from an earlier saved MATLAB data file. The interface is shown in Figure 3.

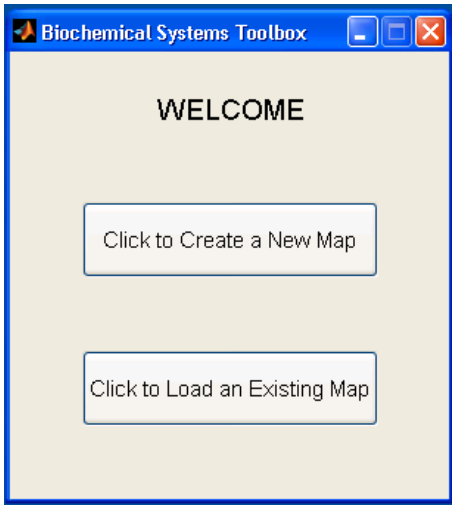


Figure 3: Opening GUI of *BSTBox* as it is invoked in MATLAB using the command 'BST'; the user may opt to create a new map or to retrieve an earlier created and stored map from a MATLAB data file.

To illustrate the software, consider the pathway in Figure 4, which describes the regulation of glycolysis in the bacterium *Lactococcus lactis*. The flow of material in this pathway is governed by the following steps (for abbreviations, see legend of Figure 4):

1. Phosphorylation of glucose into G6P, involving PEP, as governed by the phosphoenolpyruvate: carbohydrate phosphotransferase system (PEP:PTS);
2. Phosphorylation of isomerized G6P (FBP) to form FBP using ATP;

3. Cleavage of FBP into two molecules of 3-PGA involving the consumption of two molecules of inorganic phosphate (P_i) and generation of two molecules of ATP (in reality this step consists of several steps which are condensed here);
4. Reversible conversion between 3-PGA and PEP;
5. Catalysis of PEP yielding two molecules of ATP;
6. Unspecified minor consumption of PEP for use in other metabolic pathways.

In addition to the flow of material, the regulation of the pathway is of importance. In this case study, one known pair of modulators and one hypothesized inhibition process are considered. These are:

1. Investigations by Prichard [58], Mason [59], and others suggest strong activation of pyruvate kinase activity by FBP and inhibition by inorganic phosphate, P_i. These effects are to be included in the model.
2. Galazzo and Bailey [60] found significant inhibition of glucose uptake by G6P in yeast. So far this regulation has not been reported (or refuted) for *L. lactis*. *BSTBox* allows exploring this possible regulation by initially including a specific parameter for this process. If the analysis identifies the value of this parameter as zero, the effect is deemed insignificant.

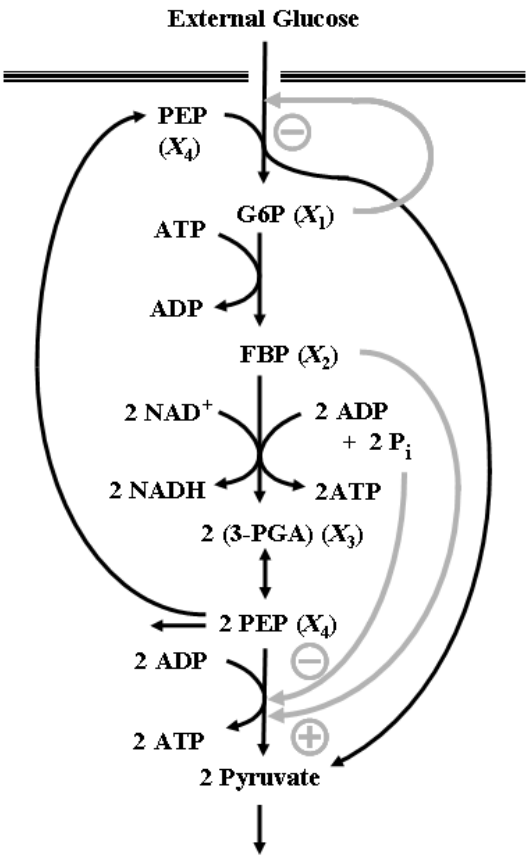


Figure 4: Simplified representation of glycolysis and lactate production in *L. lactis*. Black arrows show flow of material, while grey arrows indicate signals; subscripted *X*'s designate dependent variables in the model equations. Metabolites without symbolic names were used as offline variables (see *Text*); Pyruvate is only shown for completeness but is not explicitly modeled. Abbreviations: G6P: glucose 6-phosphate; FBP: fructose 1,6-bisphosphate; 3-PGA: 3-phosphoglycerate; PEP: phosphoenolpyruvate; ATP: adenosine triphosphate; ADP: adenosine diphosphate; P_i: inorganic phosphate

Preparing the Data. The first step of designing a new model consists of deducing symbolic equations from a list of components. For this task, it is assumed that the user has an MS-Excel spreadsheet with true experimental or alleged time course data, identified in the first row by a title for each column. *BSTBox* screens the Excel file and presents the user with the GUI shown in Figure 5, displaying five list categories, from which the user selects: 1) MS-Excel sheet-name containing raw data; 2) MS-Excel sheet-name containing smoothed data (if any); 3) column title containing time data; 4) column title(s) containing time course data for dependent variables to be modeled; and 5) column title(s) containing (constant or varying) time course data for "offline" and/or independent variables.

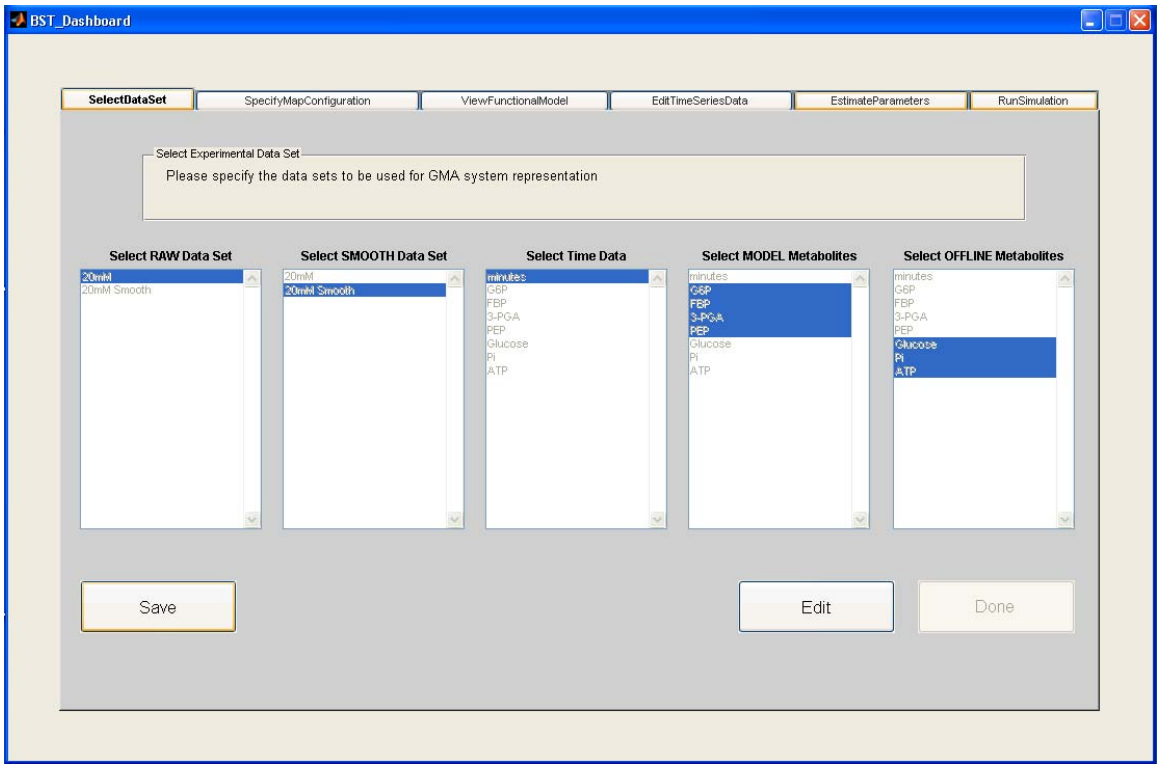


Figure 5: Different functionalities of *BSTBox* are embedded in separate tabs, which unlock progressively as the user transitions from one stage of the modeling process to the next. When creating a new biochemical map, the first tab, "Select Data Set", allows the user to specify which MS-excel sheet contains the raw and smoothed data, and also which columns contain the time data and the experimental measurements of the metabolite levels. At this step, the distinction is made between metabolites to be modeled and metabolites to be treated as offline. Upon specification, the user clicks 'Done' to proceed to the next tab.

Creating a Map. The model analysis may now proceed in different ways, depending on how much is known about the underlying pathway structure. At one extreme, the pathway structure and its regulation may be known in great detail. This is the case for the *Lactococcus* case study, with the possible, minor exception of G6P inhibiting or not inhibiting glucose uptake. In this situation, BST prescribes directly which variable is to be included in which processes, and this information is already sufficient to formulate equations [13]. At the other extreme, nothing much specific may be known about the pathway. In such a case, all variables enter all equations, and the later parameter estimation step will ideally identify which parameter values are zero in

the optimized solution, thereby indicating which affects of variables on processes are real and which ones not. This situation of structure identification has been discussed widely in the literature (*e.g.*,[56]). Not surprisingly, it is much more complex than inverse modeling with a known pathway structure.

Once all constituents of the map are identified, the user specifies, as well as contextual information allows, how these components interact and influence each other. As has been discussed many times in the literature (*e.g.*, [13]; also see above), a key distinction to be made here is between the flow of material among constituents, such as in the phosphorylation of glucose to glucose-6-phosphate, and the modulation of a process, such as the feedforward activation of pyruvate kinase by FBP. In many cases, a constituent may have several distinct roles. It may be the product of one reaction and the substrate for another reaction, and furthermore modulate one of the reactions modeled in the system. Often, this information has to be mined from experimental literature or characterized in collaboration with a subject area expert. It is often convenient to organize this information as a list of processes in the system and to associate with this list the components that contribute to or modulate those processes. Table 1 summarizes this information for the glycolytic pathway.

Table 1: Enumeration of features of the biochemical map, exemplified with the *Lactococcus* case study. The table lists, for each metabolite in the first column, the number of influxes and effluxes, and also the components affecting each of these fluxes. For instance, the concentration of 3-PGA in the system is determined by the dynamic balance between the sum of two separate influxes, one from FBP and the other from PEP, and one efflux toward PEP.

Metabolite	Influxes and their Effectors	Effluxes and their Effectors
G6P	(1) Glucose, PEP, G6P	(1) G6P, ATP
FBP	(1) G6P, ATP	(1) FBP, P_i
3-PGA	(1) FBP, P_i (2) PEP	(1) 3-PGA
PEP	(1) 3-PGA	(1) Glucose, PEP, G6P (2) PEP, FBP, P_i (3) PEP (4) PEP

As a specific example, the dynamics of PEP is affected by one production process, namely through degradation of 3-PGA, and four degradation processes, namely: 1) the PTS mechanism, which is affected by glucose, PEP and G6P; 2) catalysis by pyruvate kinase into pyruvate, modulated by FBP and P_i; 3) degradation back into 3-PGA; and 4) unspecific channeling into other metabolic pathways.

BSTBox provides a GUI (Figure 6) that permits the direct specification of all such processes of a table. It allows the user to select one metabolite at a time, in the left-most list, and then to include in the other lists the components that contribute to or modulate the influx or efflux for that particular metabolite. The buttons immediately below these lists allow the specification of the number of such fluxes as well as their inspection, one at a time. A right-click menu allows the user to add or delete metabolites to or from the list.

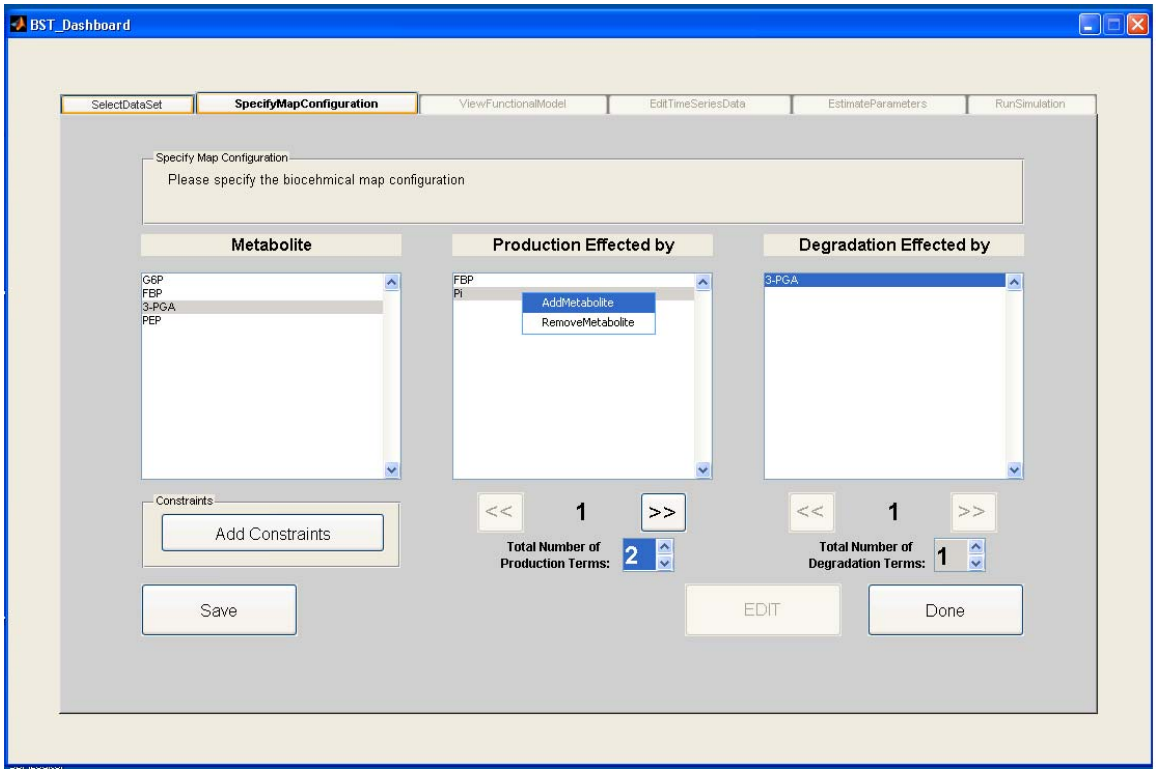


Figure 6: The second tab, 'Specify Map Configuration', allows the user to enumerate all processes that determine the dynamics of each metabolite. The user specifies the numbers of influxes and effluxes that determine the concentrations of each metabolite, and for each of these processes lists the variables that directly influence that process. In effect, the user provides a tabular description of what would otherwise be a graphical biochemical map. The user may also add constraints, such as the conservation of mass between precursors and products. The corresponding GUI for this task is invoked by clicking the button 'Add Constraints' (see Figure 7). When done, the user proceeds to the next tab to generate and view the functional form of the model equations.

Another feature is the option of accounting for constraints such as precursor-product relationships. This specific GUI is shown in Figure 7. Constraints should be implemented only after due consideration, as was discussed extensively in [61]. As an example for possibly unexpected complications, consider the unbranched, irreversible reaction step between G6P and FBP (Figure 4). According to the map, all material leaving the G6P pool immediately enters the FBP pool, suggesting a "hard" precursor-product constraint. However, it is known that organisms have secondary routes of generating and using G6P, which are not included in the model but are presumably active in the organism and affect the observed time courses. Thus, insistence on hard constraints might sometimes be too restrictive. On the other hand, omission of known constraints increases the parameter search space, which is often a disadvantage. No general guidelines can be given, except that cautious consideration is advised.

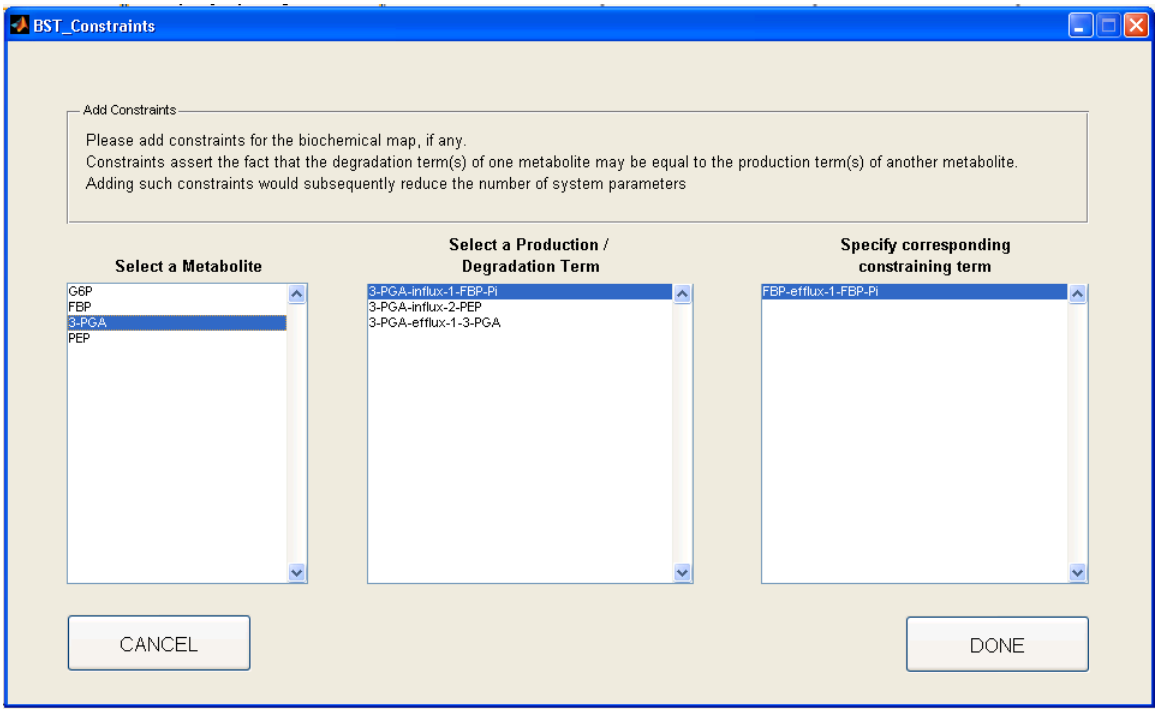


Figure 7: This GUI permits specification of precursor-product relationships between different influx and efflux terms. For each metabolite, the user may browse the list of processes (influxes and effluxes) that determine the levels of that metabolite. Given an unbranched pathway, such as in the case of *Lactococcus*, the user may constrain the efflux from one pool to be equal to the influx into another pool. Currently *BSTBox* supports only one-to-one constraint specifications.

Formulating Symbolic Equations. Given the list of fluxes and their effectors, *BSTBox* permits the formulation of GMA or S-system equations with a few clicks. Figure 8 shows the GUI for generating the (not yet parameterized) model system, based on the table of processes and constraints that were specified earlier. *BSTBox* also offers the option to view and format this system of equations in multiple formats like HTML, MS Word and MS PowerPoint. For instance, the equations for the *Lactococcus* case study, formatted in HTML, are:

$$\dot{X}_1 = \alpha_1 X_1^{g_{11}} X_4^{g_{14}} \text{Glucose}^{g_{15}} - \beta_1 X_1^{h_{11}} ATP^{h_{1,ATP}}$$

$$\dot{X}_2 = \beta_1 X_1^{h_{11}} ATP^{h_{1,ATP}} - \beta_2 X_2^{h_{22}} P_i^{h_{2,Pi}}$$

$$\dot{X}_3 = 2\beta_2 X_2^{h_{22}} P_i^{h_{2,Pi}} + \alpha_{3a} X_4^{g_{34}} - \beta_3 X_3^{h_{33}} \qquad \text{..... Eq. 4}$$

$$\dot{X}_4 = \beta_3 X_3^{h_{33}} - \alpha_1 X_1^{g_{11}} X_4^{g_{14}} \text{Glucose}^{g_{15}}$$

$$- \beta_{4a} X_2^{h_{42}} X_4^{h_{44a}} P_i^{h_{4,Pi}} - \alpha_{3a} X_4^{g_{34}} - \beta_{4b} X_4^{h_{44b}}$$

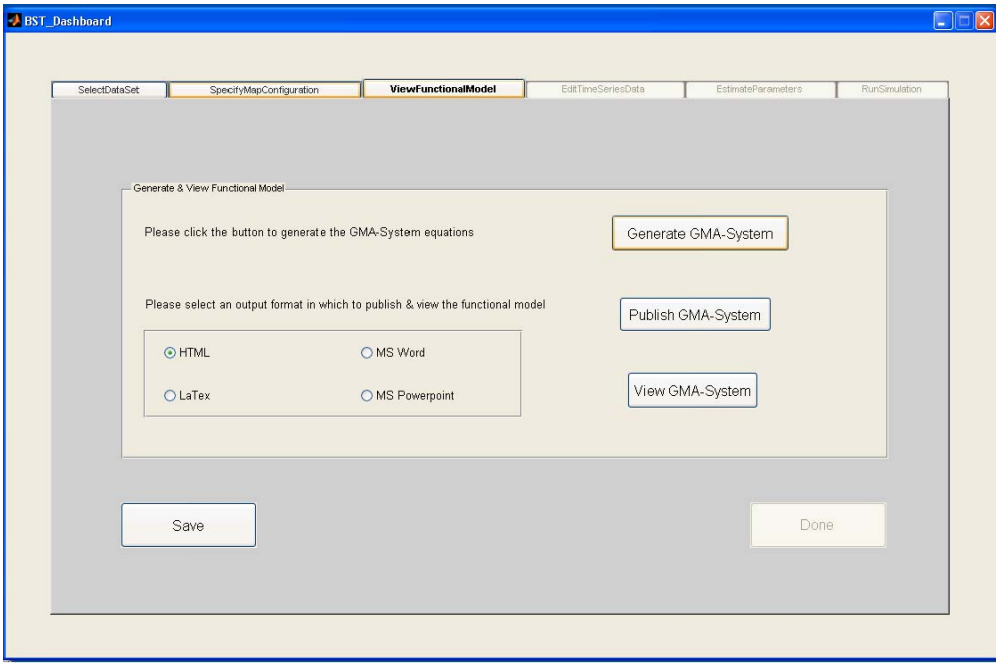


Figure 8: The tab 'View Functional Model' allows the user to generate the necessary system of BST equations by the simple click of a button. Due to the rigorous rules of BST, *BSTBox* has enough information in the lists of processes and components to generate a (not yet parameterized) model and to present the model equations in a variety of formats. When done, the user proceeds to the next tab to specify or edit the time course data interactively for all metabolites involved.

At this juncture, the user has achieved only a partial specification of the model system. It is partial since the parameter values are still unknown or unspecified. If the parameter values are unknown, which is usually the case, the user may use the "*Estimate Parameters*" tab, as detailed in a later section ahead. If the parameter values are known, the user may bypass the parameter estimation phase by directly typing these values into the GUI as shown in Figure 13.

Editing Time Course Data. BSTBox allows the user to visualize and edit the data on which the model is to be based. Data manipulations are initiated with the "*Edit Time Series Data*" button (shown on the GUI in Figure 9), which in turn invokes the *SplineTool* (shown in Figure 10). The *SplineTool* allows the user to smooth time courses with predefined filters or with curve-fits such as cubic-spline interpolants or some least-squares approximation. The user may also manually add or delete time points to the curve. An added feature allows the user literally to move a given point into a desired position. This is accomplished by increasing or decreasing the *x*- or *y*- value of a data point using the '+' / '-' buttons on the lower left section of the *SplineTool* GUI. After the completion of data entering and editing, the user has the option to export the data into an MS-Excel file from the *BSTBox* GUI (Figure 9). Permitting this type of inclusion of hypothetical data opens an entirely new realm of modeling possibilities, as will be exemplified later.

For the *Lactococcus* case study, actual time course data characterizing key metabolites of the pathway are available in the form of *in vivo* NMR measurements of all involved variables [34, 51, 52]. As explained above, the *BSTBox* GUI (Figure 9) in combination with the *SplineTool* (Figure 10) allows the user to view the metabolic time series data, to edit and smooth them, for instance with a spline, and to toggle between raw and smoothed data. Smoothing is often computationally beneficial for purposes of parameter estimation. If there is suspicion that the smoothing process might introduce undue bias, the raw data may be fitted again, once the estimation based on the smoothed

data has produced reasonable initial parameter estimates. Using the *SplineTool*, each experimental time course for the *Lactococcus* case study was edited, by manually removing noisy data points. The semi-smoothed curves, thus obtained, are shown below in Figure 11.

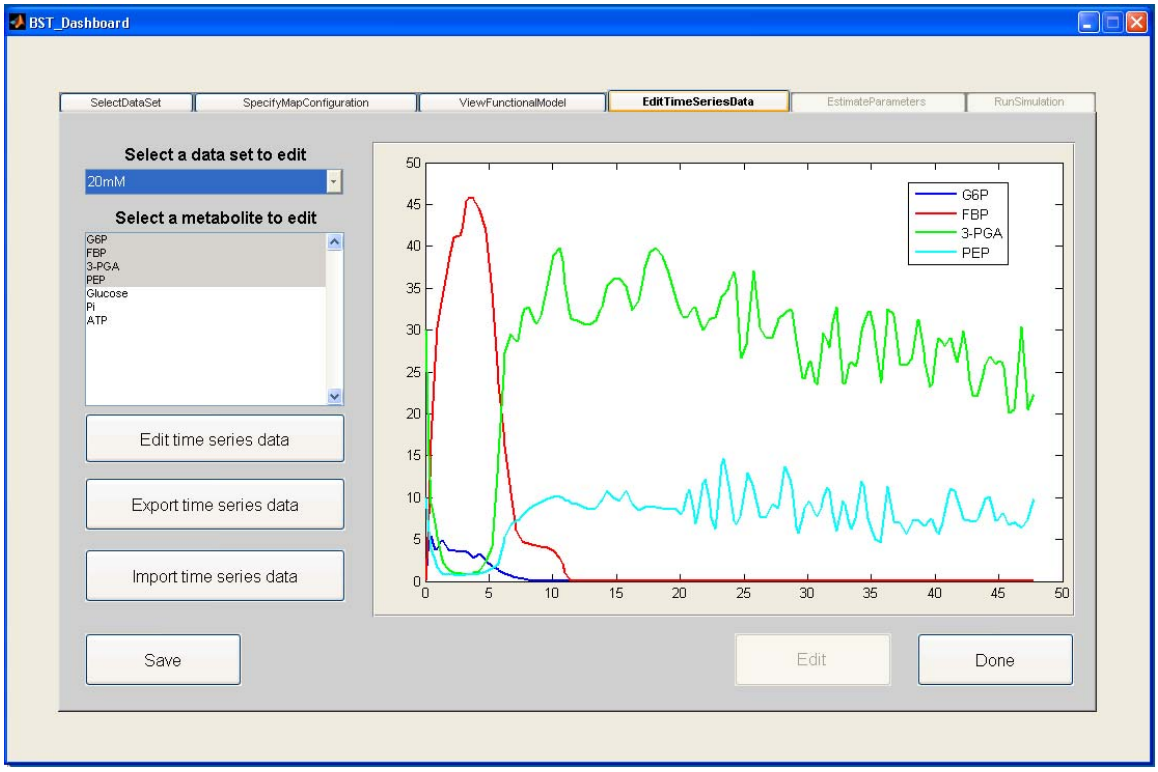


Figure 9: The tab 'Edit Time Series Data' allows the user to view the raw experimental time course data and to smooth them; it is easy to toggle between the raw or smooth data sets by means of a drop down menu. The user may select multiple metabolites to view their time courses simultaneously. When editing these time plots, *BSTBox* invokes MATLAB's Spline Toolbox, shown in Figure 10, to allow the user to edit and approximate these time curves using cubic splines. In the absence of experimental observations, users of *BSTBox* have the ability to start with time courses (in MS-Excel) that had not been measured but are expected based on the biologist's experience and intuition, load these data into *BSTBox*, smooth them, specify a model, and test hypotheses with this model. The user has the option to export the edited time series data as an MS-Excel file. As a future enhancement, the user will be allowed to import and update these data sets with additional experimental observations. When done, the user proceeds to the next tab to estimate or specify values for the system parameters.

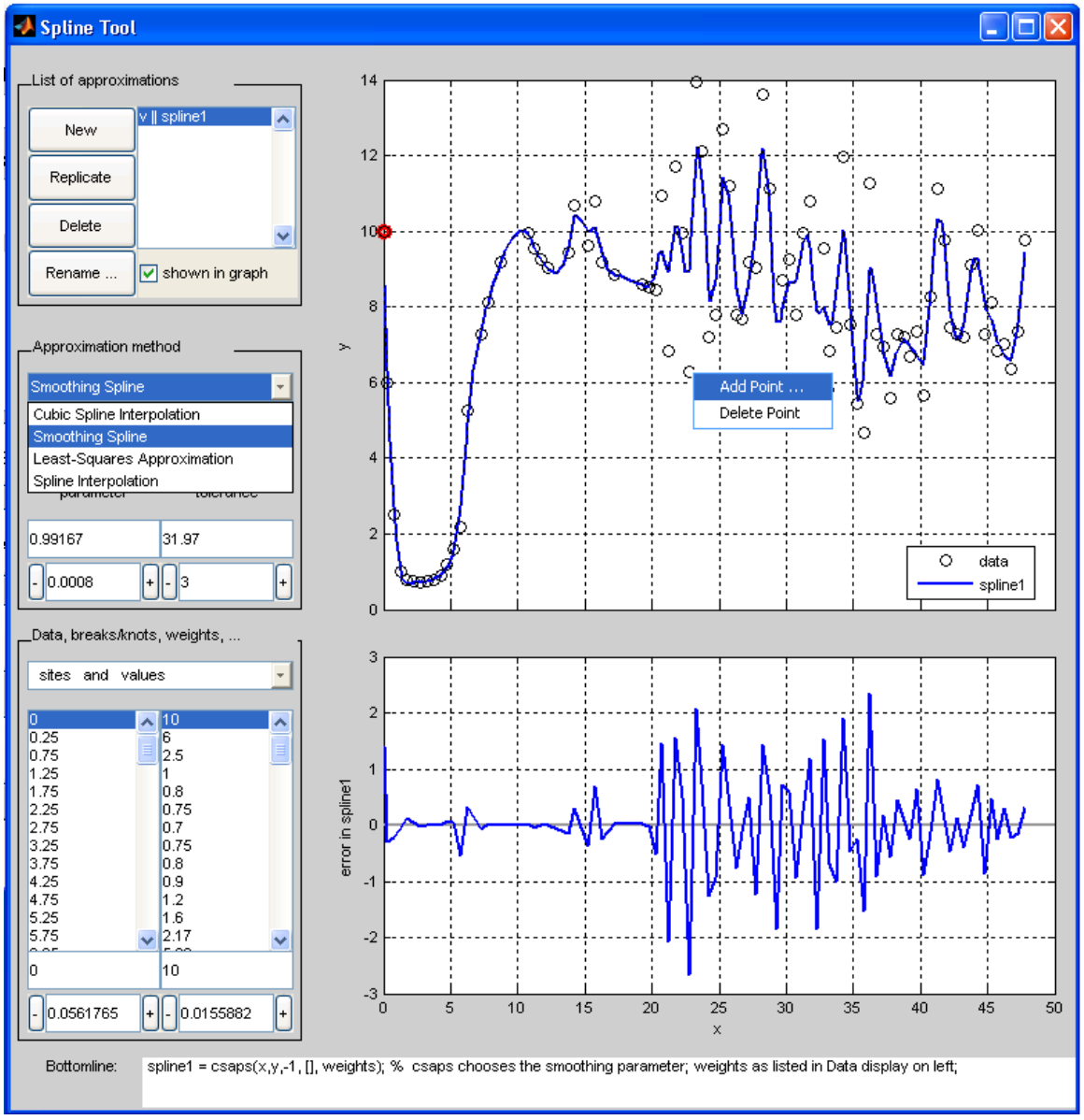


Figure 10: SplineTool for editing time series data. *BSTBox* provides access to this MATLAB tool such that the user's time series data are directly loaded into the Spline Tool interface and made available for editing purposes. The user may add or delete time points using the right click menu in the upper plot area; the lower plot shows the error in the approximated curve from the original (raw) time series data; the GUI elements on the left hand allow the user to choose between multiple approximation schemes such a cubic spline interpolation or least-squares approximation; on the lower left hand side, the user may manually edit each data point in the table of values listed here. All changes are preserved and saved in *BSTBox*.

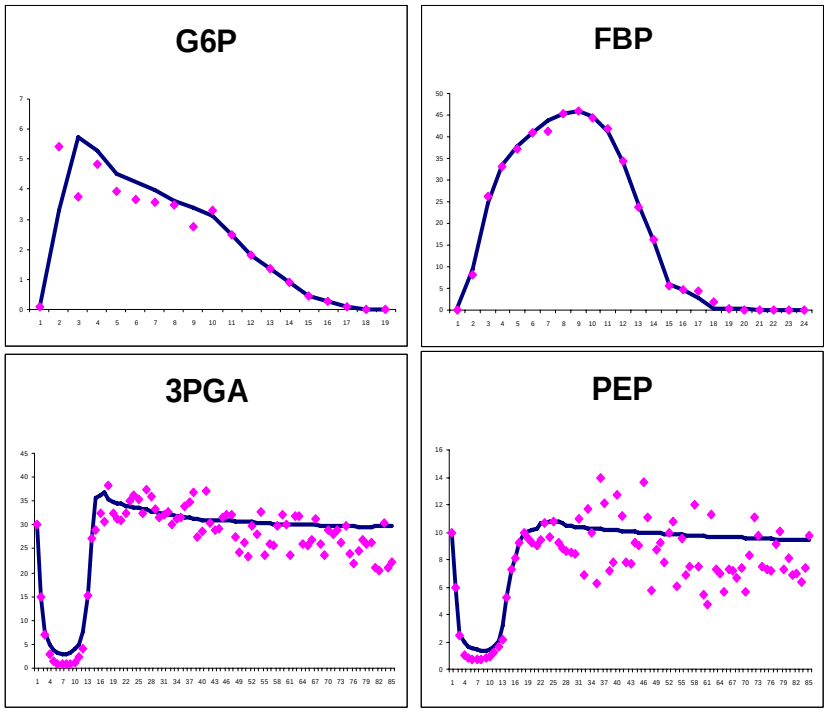


Figure 11: Smoothed Time Course Data; each of these curves was manually enhanced, using the Spline Tool interface. A cubic spline interpolation was used and data points were manually deleted from the set of points available, generating the curves shown here. magenta: raw data; blue: manually smooth data

For concept modeling, it is important to have the facility to enter and manipulate hypothetical time courses that had not been measured but are based on the biologist's intuition. An example could be that, after a particular stimulus, variable X_3 is expected to rise sharply in a sigmoidal fashion to about twice its baseline value, even though specific data are not (yet) at hand. In this case, the user could start with an empty spreadsheet and create data according to his or her qualitative knowledge about the data. For instance, suppose variable X_3 is initially assumed to be at its nominal value. A data point $X_3(t=0)$ at time zero is created with the known nominal value or with a value of 1, which would correspond to a representation of the variable in a scaled manner. If doubling of the variable occurs within twenty minutes, the user specifies the corresponding value of

$X_3(t=20)$ and have the option of creating hypothetical data connecting the two points in a sigmoidal fashion and to refine them, where necessary, with the manual editing tool.

Parameter Estimation. Given time course data and a mathematical model in the form of a system of differential equations, the estimation of parameter values constitutes an inverse problem. Many methods of optimization and system identification have been developed throughout the past decades. Notable examples include nonlinear regression [21, 62, 63], genetic algorithms [23, 64, 65], evolutionary programming [66, 67], and simulated annealing [25]. Sadly, none of these methods is ideal, and parameter estimation continues to be the bottleneck of mathematical modeling.

Conceptually, the process of parameter estimation involves: (1) construction of an *objective function* such as the sum-of-squared-errors between model and data; (2) selection of an initial guess for each parameter; and (3) application of a numerical, iterative optimization scheme designed to minimize the objective function. This process is depicted graphically in Figure 12.

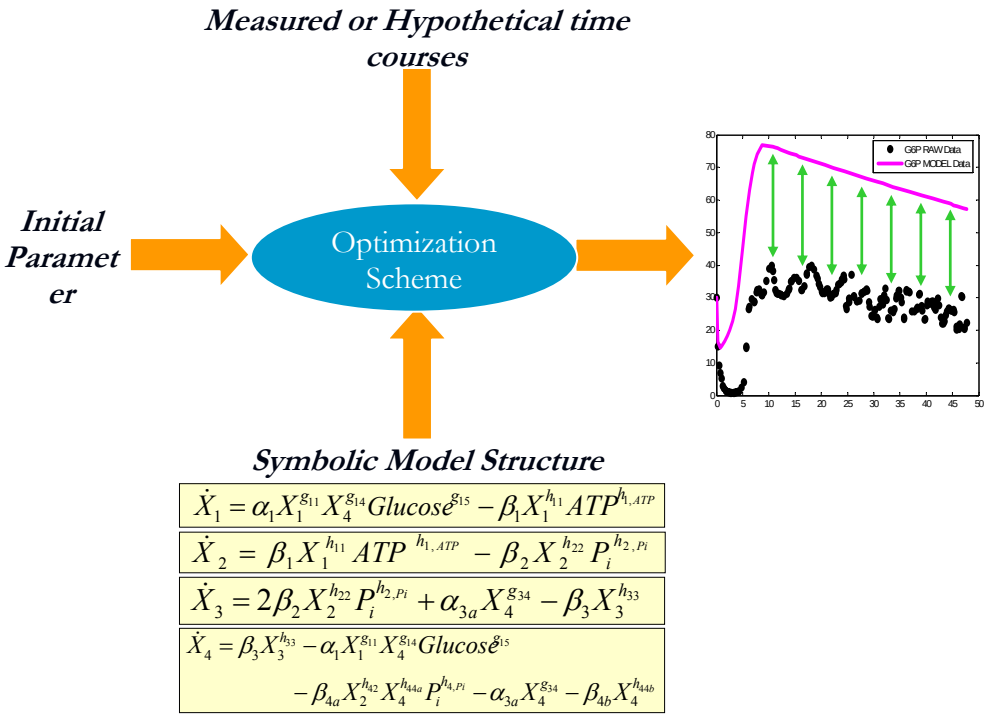


Figure 12: General architecture of an optimization scheme for parameter estimation. Three sets of "inputs" are the measured or hypothetical data, the symbolic (not-yet parameterized) model equations, and initial guesses for all parameter values. The optimization algorithm begins by solving the equations with the guessed parameter values and computing the residual error between the solution and the data. It then determines new parameter guesses, solves again, and computes the error again. This cycle is iterated thousands of times, in an attempt to minimize the error between the computed model time courses and the data curves ("shrinking" of the green arrows, indicating that the residual error is to be minimized).

The enormous technical difficulties in executing this seemingly straightforward process have been explored in great detail in previous research on *Lactococcus* case study [68, 69]. A diverse repertoire of techniques were explored and analyzed, including local gradient-based techniques such as regular-ODE based estimation and slope-based estimation, as well as global optimization techniques like genetic algorithms. Many of these techniques are directly accessible from the *BSTBox* GUI illustrated in Figure 13, which also allows the user to specify technical features like upper and lower bounds for parameter values, initial guesses, an acceptable residual error, and the choice of a numerical ODE solver. The user may choose to estimate parameters while minimizing the error with respect to the raw experimental data or a smoothed data set, and also select the range of time over which each parameter is to be optimized. Results associated with the *Lactococcus* case study, obtained with *BSTBox*, are shown in Figure 14.

Simulation and Analysis. Once the system has been fully identified, complete with all variables, processes, and parameter values, the *BSTBox* GUI shown in Figure 13 allows the user to simulate, by the click of a button, the entire system, a single component, or a subset of components. As during parameter estimation, if only a sub-system is being simulated, *BSTBox* automatically feeds the system values for the "offline" components upon cubic spline approximation. The GUI then displays the results from the simulation together with the experimental data, which facilitates comparisons and assessments of quality of fit. These simulation results may also be viewed in separate windows.

One of the common analyses with computational models is the investigation of changes in system dynamics in response to changes in the initial values of the system. Known as "perturbation studies", such analyses entail examining whether the system returns to its original or a different steady state or possibly to a different attractor like a stable limit cycle. To facilitate such analyses, the *BSTBox* GUI provides a placeholder for specifying different initial conditions for each dependent variable.

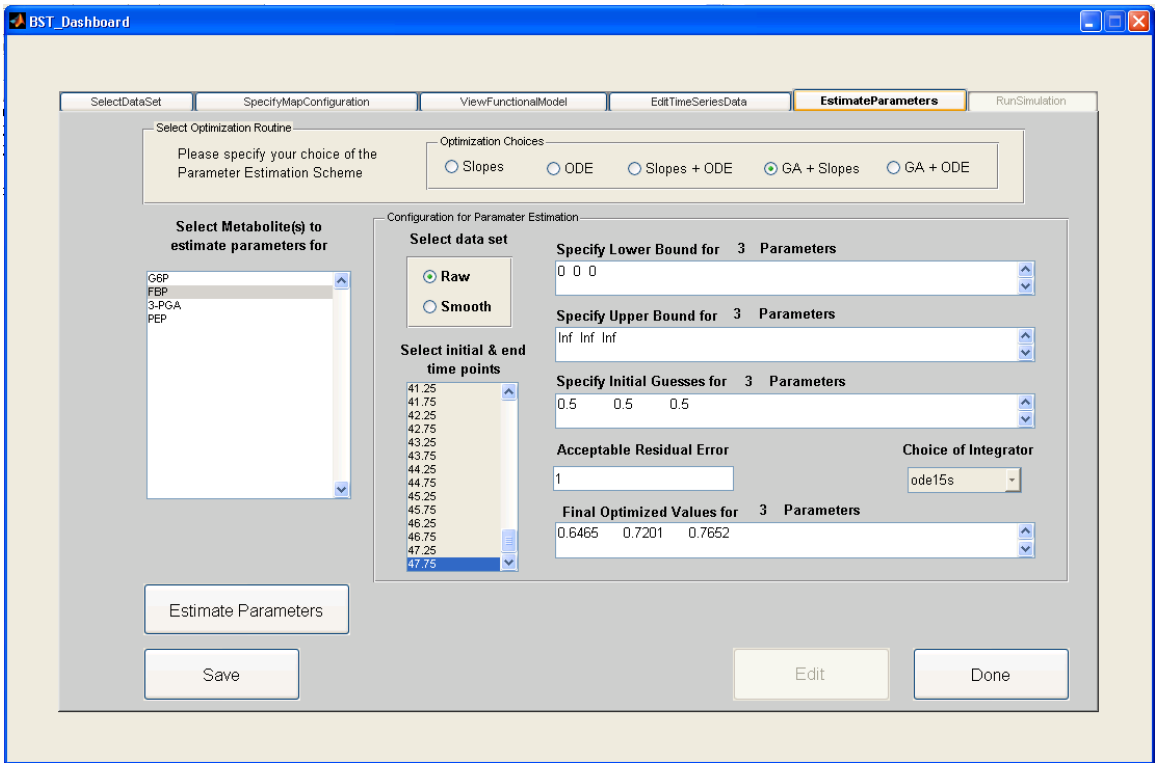


Figure 13: The 'Estimate Parameters' tab provides an interface to call various parameter estimation techniques. From the radio button selection in the top panel, a user can select either slope-based and/or ODE-based local optimization, or a global optimization technique such as a genetic algorithm. The user may change the default settings for the bounds of parameter search, initial parameter guess and acceptable residual error. When using ODE-based estimation, the user may also specify the desired numerical solver to be employed for integrating the equations. The user furthermore chooses whether to search for parameters that would make the underlying model fit the raw data or the smoothed data and determines the time period of the data over which the parameters should be optimized. Building upon the offline-spline-based approximation framework, the user can estimate parameters for either the entire system simultaneously, for one parameter at a time, or for combination(s) thereof. Finally, the user has the option to 'bypass' parameter estimation if the parameters are already available. The user can directly type these parameter values into the GUI and proceed to the next tab to simulate the system. In such a scenario, the user does not have to provide time series data, and only initial values at the start time are expected in the MS-excel file.

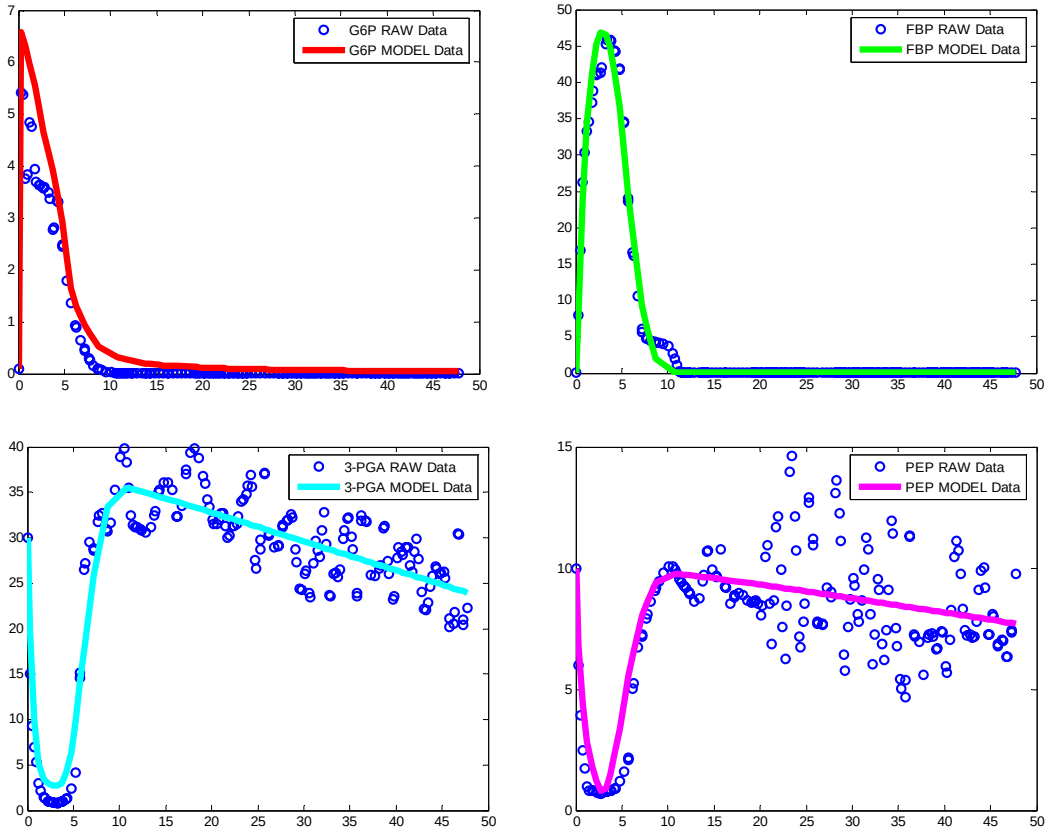


Figure 14: Model fits (lines) to the *Lactococcus* data (symbols), obtained with parameters estimated using regular ODE-based optimization scheme. The initial parameter guesses were set to +0.5 or to -0.5, depending on the sign of each parameter, which is usually known from the tenets of BST. The lqscurvefit routine, available within the MATLAB Optimization Toolbox, was used. The complete optimization took about 20 minutes. Note that the raw data for the initial five minutes, for both 3-PGA and PEP, were made-up i.e. they are artificial data points. In reality, PEP and 3PGA are not detected before addition of labeled glucose, because they are unlabeled, but after the glucose bolus and while glucose is present they are not detected because their concentration is below the detection limit.

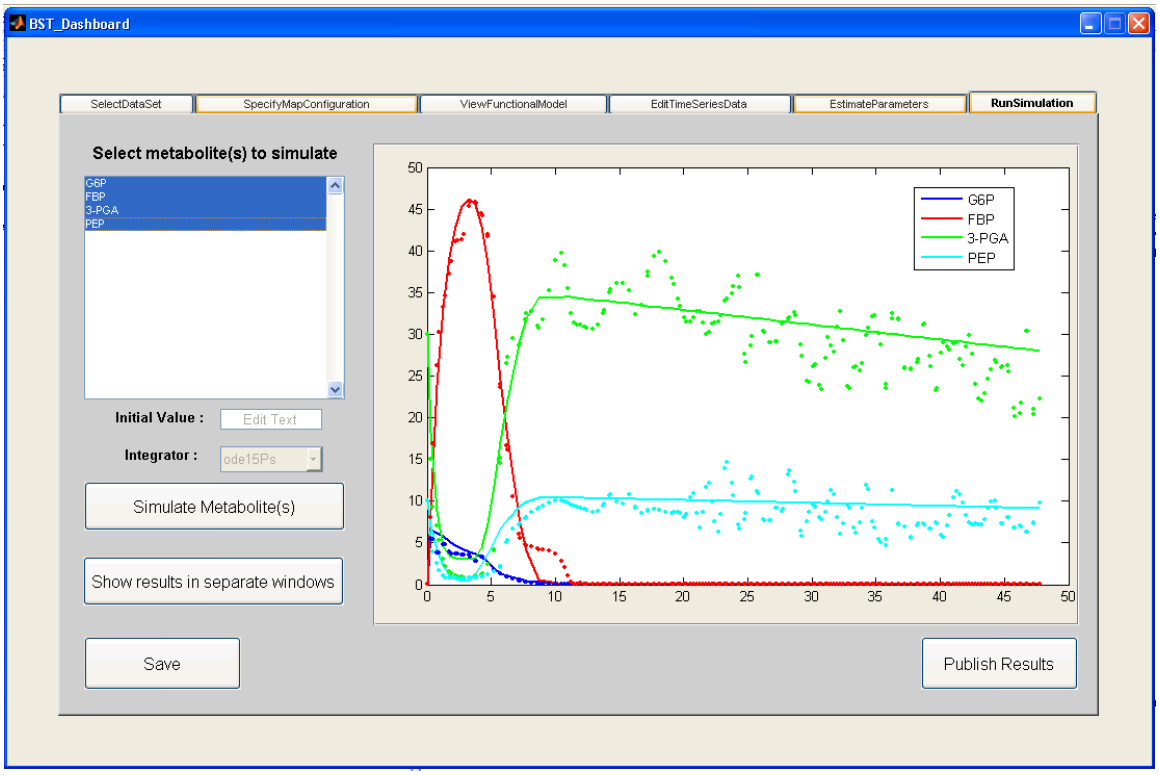


Figure 15: The tab 'Run Simulation' uses the functional model, combined with initial time values and final parameter values from previous tabs to simulate the system. The user may choose to simulate either the entire system, one component at a time, or a subset of components. As earlier, spline approximations are used for all offline components. For assessments of quality, the GUI displays the results together with the data or in separate windows. It is easy with this interface to conduct 'perturbation studies' by changing the initial values for one or some of the metabolites. The user can also choose different integration schemes to simulate the system. Upon the successful execution of the simulation, the user may export the simulation results, which include all categories of time series data – raw, smoothed/edited and simulated, to an MS-Excel file.

Concluding Remarks

In the past, the primary role of mathematics in molecular biology has been bookkeeping, first as a means of recording and storing quantitative information, and more recently, with the advent of bioinformatics, as a means of mining and interpreting the enormous amounts of data generated by high-throughput methods. Greatly increased computer power, advanced mathematical techniques, and the availability of data of vastly enhanced quality and quantity are now slowly beginning to move mathematics into the more prominent role of integrating information, offering predictions, and guiding future experimentation through the generation of computationally inspired hypotheses. A crucial component of this emerging "computational systems biology" is the development of mathematical models. These have traditionally been constructed from the bottom up, that is, by assembling network models from representations of local features. More recently, and complementing traditional approaches, effort has been dedicated to top-down model development and parameterization from time series data, which modern methods of molecular biology are generating with increasing frequency. In this chapter, a novel method is proposed that bridges the gap between semi-quantitative biological knowledge and the construction of detailed mathematical models. The starting point for this method is a concept map, which shows connections and interactions between components of biological systems and responses. This type of map is very prevalent in the biological literature, yet it has not really been exploited for modeling purposes. The key toward model construction is the translation of biological expertise, experience, qualitative insights and intuition associated with the map into quantifiable temporal responses of all components. This translation subsequently allows the application of modern inverse methods for the determination of parameter values that specify the model and render it useful for analysis and simulation. The method of concept map modeling thus has the potential of converting biological insight into a concrete mathematical model that may be used to test assumptions and generate testable hypotheses.

CHAPTER 3

DYNAMIC FLUX ESTIMATION (DFE) [2]

At the center of computational systems biology are mathematical models translated from concept maps (as described in the previous chapter) that capture the dynamics of biological systems and offer novel insights. While there exist several software tools to assist in this translation, the bottleneck in the construction of these models is presently the identification of model parameters that make the model consistent with observed data. Dynamic Flux Estimation (DFE) is a novel methodological framework for estimating parameters for models of metabolic systems from time series data. DFE consists of two distinct phases, an entirely model-free and assumption-free data analysis and a model-based mathematical characterization of process representations. The model-free phase reveals inconsistencies within the data, and between data and the alleged system topology, while the model-based phase allows quantitative diagnostics of whether—or to what degree—the assumed mathematical formulations are appropriate or in need of improvement. Hallmarks of DFE are the facility to: diagnose data and model consistency; circumvent undue compensation of errors; determine functional representations of fluxes uncontaminated by errors in other fluxes; and pinpoint sources of remaining errors. The results presented here suggest that the proposed approach is more effective and robust than presently available methods for deriving metabolic models from time series data. Its avoidance of error compensation among process descriptions promises significantly improved extrapolability toward new data or experimental conditions. This chapter outlines the theory and application of DFE

[2] Part of this chapter is published in: G. GOEL, I.-C. CHOU and E. O. VOIT, "System estimation from metabolic time-series data". Bioinformatics. 24(21): p. 2505-2511, 2008

and demonstrates how this technique is useful in circumventing some key issues in parameter and system estimation from time-series data.

System estimation from time-series data

Recent advances in molecular and systems biology have provided us with a strikingly novel parameter estimation strategy, which is based on experimentally determined time series of observations at the genomic, proteomic, or metabolic levels. These time profiles contain enormous information about the structure, dynamics and regulatory mechanisms that govern the biological systems of interest. However, extraction and integration of this information into fully functional, explanatory models is a daunting task, and about one hundred articles have appeared within the past ten years, each improving certain aspects of the estimation process. Most of them used regression, genetic algorithms, simulated annealing, or different evolutionary approaches [65, 67, 70-75] to attack the main problem of optimizing parameter values against the observed time series data. Other papers developed support algorithms, for instance, for smoothing overly noisy data, characterizing basins of attractions containing solutions with minimal error, or circumventing the costly integration of differential equations [29, 31, 50, 74, 76-79].

All of the proposed estimation methods developed up to date face significant problems in four distinctly different classes:

1. ***Computational issues, including:*** slow algorithmic progress toward the error minimum or lack of convergence; very complicated error surfaces with numerous local minima; substantial time requirements for integration of differential equations.
2. ***Data related issues, including:*** overly noisy data; missing data; missing time series; collinearity between time series; solution spaces with equal error; non-informative, *e.g.*, essentially constant, time profiles.

3. ***Mathematical issues, including:*** distinctly different, yet equivalent solutions; non-equivalent solutions with similar error; invalid assumptions regarding the chosen process descriptions; error compensation within and among flux descriptions and within and among equations (see illustrations in the next section).
4. ***Issues of model quality beyond goodness of fit, including:*** lack of diagnostic tools beyond the residual error; lack of model fit for data not used in the estimation; model failure in extrapolations; lack of criteria for optimality of the obtained parameters; lack of criteria for determining the appropriateness of the chosen mathematical representations; lack of methods for assessing whether residual errors are due to idiosyncrasies or noise in the data, an invalid model structure, inadequate computational methods, or a combination thereof.

Many articles have acknowledged and discussed various computational issues in great detail and some have addressed issues related to data and models. However, there has been little if any substantial discussion of model validity and quality beyond residual errors, except for the common statement that the estimated parameter set may not be unique.

Issues of error compensation

All parameter estimation algorithms encounter the risk of compensating an inaccurately determined parameter value through adjusted values in other parameters. A very simple, yet instructive case occurs when two parameters p_1 and p_2 always appear as the product $p_1 \cdot p_2$. Even if p_1 is vastly under- or over-estimated, there is no residual error if p_2 is correspondingly adjusted. As an extension, systems may allow "conserved quantities" that do not change, no matter what the dynamics of the system is. The analysis of Hamiltonian systems and Lie transformation groups assesses these exact

conservation laws. In the context of BST, certain products of power-law functions may remain invariant under the action of the dynamics systems [12, 80].

Besides mathematically exact invariants, different parameter implementations of the same model structure, obtained from fitting (noisy) data, may be indistinguishable within the magnitude of the data noise. As a consequence, these different data sets are "equally good," but only with respect to the one data set used for estimation. If new data are considered, the two model implementations are likely to incur different errors. These issues are illustrate below with a didactic example.

Consider the simple biochemical pathway shown in Figure 16, which consists of a linear flow pattern, which is regulated through inhibitory and positive feedback. Specifically, the degradation of X_1 into X_2 is inhibited by the downstream product X_3, while the conversion of X_2 to X_3 is activated by X_3 itself. Equations governing the dynamics of the above system could be described with a variety of functions. For the purpose of illustration, one Michaelis-Menten process and two power-law representations are used (Eq. 5). The example demonstrates different types of redundancies that pose the risk of error compensation in tasks of estimating parameters from data.

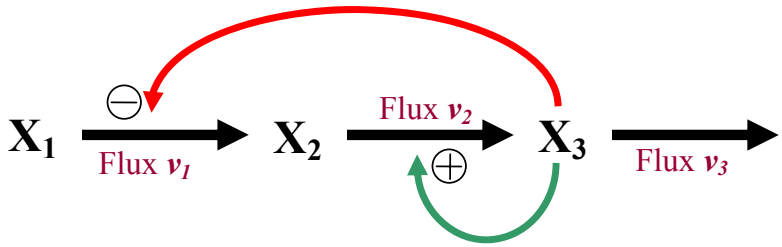


Figure 16: Simple linear pathway with feedback inhibition and one activating signal

$$X_1 = \text{Constant}$$

$$\dot{X}_2 = \frac{(X_1) * V_{\max}}{K_m\left[1 + \frac{X_3}{K_i}\right] + X_1} - p_1 X_2^{p_2} X_3^{p_3} \qquad \text{..... Eq. 5}$$

$$\dot{X}_3 = p_1 X_2^{p_2} X_3^{p_3} - p_4 X_3^{p_5}$$

Using a standard gradient-based search algorithm, such as fmincon or lsqcurvefit in MATLAB, one can easily determine alternative sets of "valid" parameters that match a set of fictitious data. Three different scenarios were analyzed whereby using lsqcurvefit different combinations of parameter values were obtained for first, flux v_1 only (in Scenario 1), secondly for flux v_1 and v_2 combined (in Scenario 2) and lastly for flux v_1 and v_3 combined (in Scenario 3). The parameters were optimized against 5,000 artificial data points in noise-free time series for X_2 and X_3.

For each scenario, the parameter values and the combined squared 2-norms of the residuals for X_2 and X_3 are listed in the respective tables below. The following general observations were made from evaluating the quantitative and qualitative information in the tables and the corresponding graphs:

A) The error residual alone is by no means a reliable indicator that the fitting parameters are numerically correct. This is clearly evident as all parameter sets yield more or less the same error residual but different, though acceptable fits, as is evident in the plots.

B) Combinations of different parameter values within the same flux term can compensate amongst each other to yield output that is mathematically not truly equivalent, but nevertheless very similar (Scenario 1; see Table 2 and Figure 17). More generally, the output can be compensated within and between fluxes, within equations, and throughout the entire system, and still produce dynamic profiles for metabolites within some error tolerance (Scenarios 2 (Table 3; Figure 18) and 3(Table 4; Figure 19)). Among these results, the underlying flux profiles can be strikingly different, but this "redundancy" would go undetected if the true internal flux profiles were not known *a priori*. It is easy to imagine that using a "wrong" set of flux

representations would lead to problems in cases of new data or other extrapolations (see next section).

Table 2: Error compensation within the same flux (v_1)

Set	V_{max}	K_m	K_i	p_1	p_2	p_3	p_4	p_5	*Residual*
1	88.2533	91.2397	1.8482	1	0.5	1	1	0.5	**6.3238**
2	18.6819	9.7831	0.5992	1	0.5	1	1	0.5	**2.0628**
3	63.0698	66.1785	1.9714	1	0.5	1	1	0.5	**7.0341**
4	91.0532	94.3597	1.855	1	0.5	1	1	0.5	**6.4499**
5	14.2804	10	1.019	1	0.5	1	1	0.5	**3.8237**
6	82.7704	87.9852	2.0162	1	0.5	1	1	0.5	**7.3094**
7	88.7362	93.0726	1.9447	1	0.5	1	1	0.5	**6.6048**
8	92.4504	97.0702	1.9466	1	0.5	1	1	0.5	**6.616**
9	68.9295	67.7172	1.6343	1	0.5	1	1	0.5	**4.9066**
10	18.2178	8.9871	0.5458	1	0.5	1	1	0.5	**2.2876**

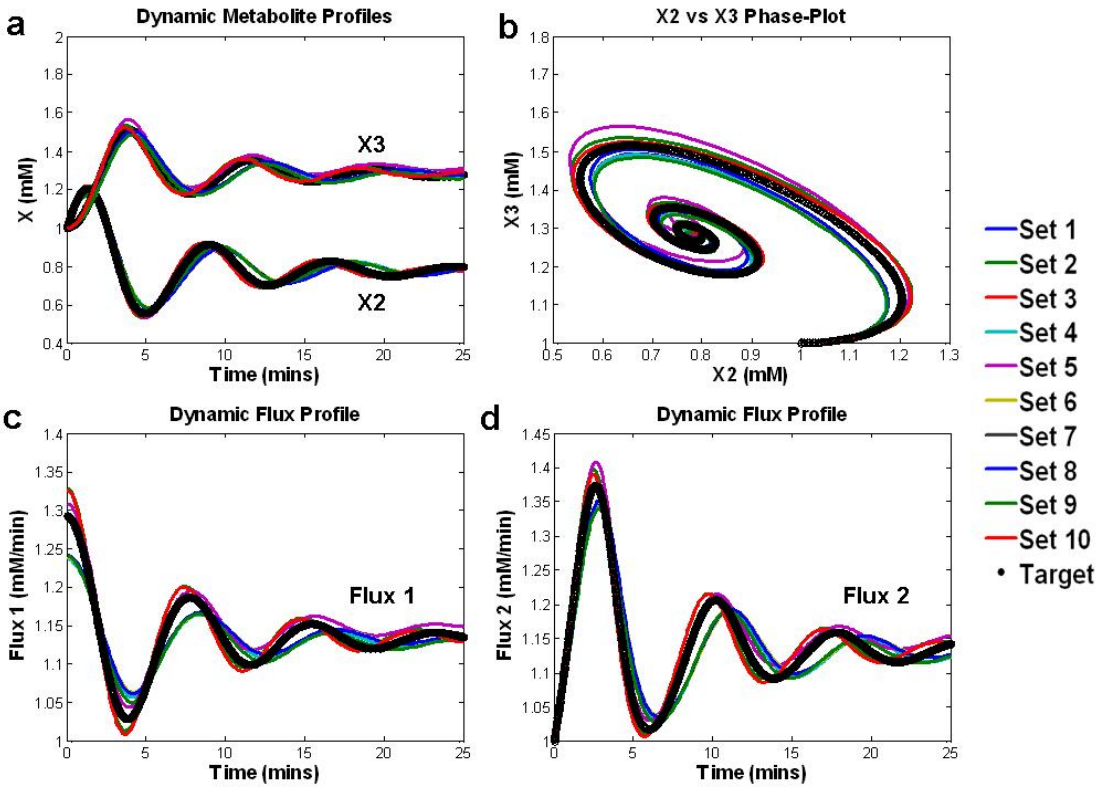


Figure 17: Although the parameters vary quite noticeably (Table 2), the residual errors do not differ much, and the resulting dynamics would hardly be distinguishable if noisy data were to be fitted.

Table 3: Error compensation between fluxes v1 and v2

Set	V_{max}	K_m	K_i	p_1	p_2	p_3	p_4	p_5	*Residual*
1	104.9701	92.1829	1.3281	1.0021	0.5785	1.0038	1	0.5	**3.4688**
2	57.0719	91.5615	15.2508	0.9401	0.9865	1.7386	1	0.5	**4.4663**
3	13.0088	9.5706	1.0968	1.0173	0.5921	0.9671	1	0.5	**6.6559**
4	103.6876	93.837	1.3967	0.9688	0.6418	1.2038	1	0.5	**5.6134**
5	12.4525	9.971	1.2927	1.0055	0.5812	1.0271	1	0.5	**2.8754**
6	10.01	8.8733	1.7075	1	0.6676	1.1052	1	0.5	**6.624**
7	124.476	88.9055	0.8893	0.9841	0.544	1.0853	1	0.5	**3.0074**
8	13.5262	9.5896	1.0152	1.013	0.6045	1.0017	1	0.5	**7.2336**
9	60.7643	96.3775	13.346	0.9117	1.0602	1.8375	1	0.5	**6.3344**
10	12.3914	9.5007	1.1869	1.0086	0.5676	1.0079	1	0.5	**2.7299**

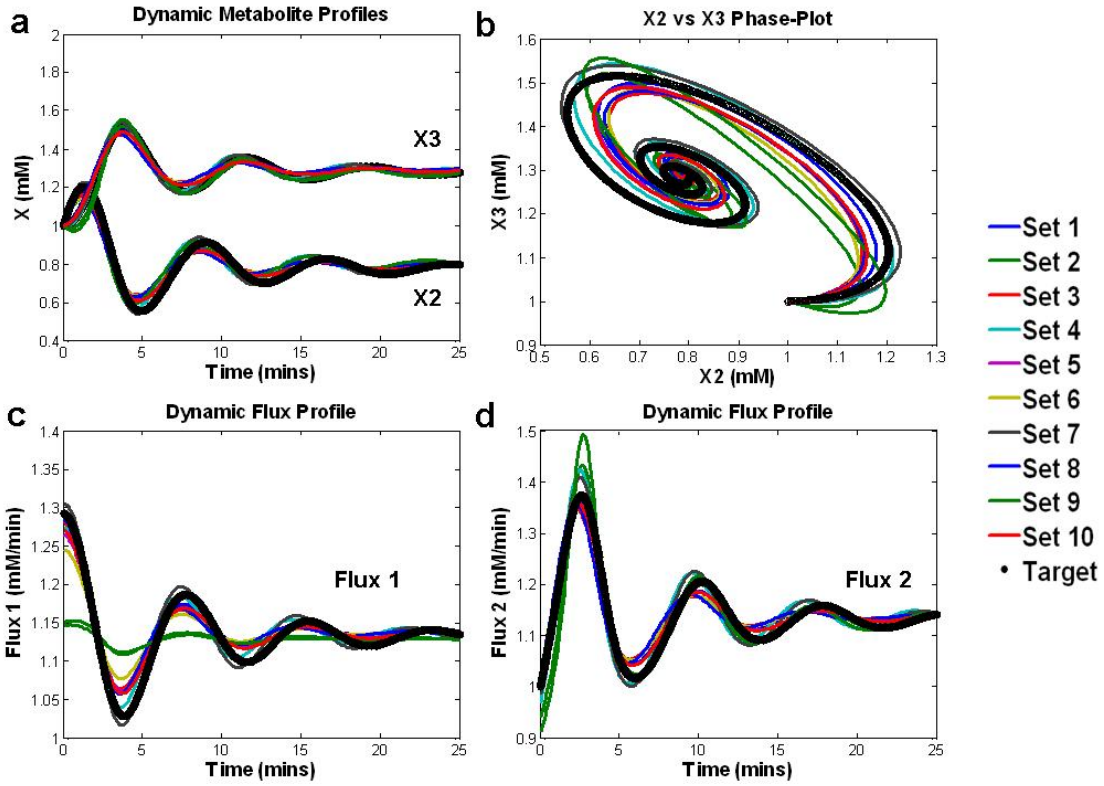


Figure 18: As in figure 17, different sets of parameter values (Table 3) may lead to similar residual errors and data fits.

Similar to the case of error compensation within a flux, error may be compensated among different fluxes within the same equation (Figure 18). As a consequence, if one flux is fitted with a "wrong" model, some or all of the other fluxes will adjust to compensate for that error and will therefore be modeled wrongly as well.

Table 4: Error compensation between different equations x2 and x3

Set	V_{max}	K_m	K_i	p_1	p_2	p_3	p_4	p_5	*Residual*
1	17.5775	9.9988	0.6979	1	0.5	1	1.001	0.5786	**4.5287**
2	19.0012	9.0003	0.5203	1	0.5	1	1.0178	0.4659	**3.2879**
3	11.0985	7.5279	1.0842	1	0.5	1	1.0001	0.5842	**7.1035**
4	16.5287	7.7719	0.5241	1	0.5	1	1.0205	0.4605	**3.5256**
5	17.8896	9.2186	0.5967	1	0.5	1	1.0206	0.4705	**3.1041**
6	87.5991	94.1804	2.1613	1	0.5	1	0.9669	0.6658	**5.1819**
7	15.5174	7.7989	0.5839	1	0.5	1	1.0011	0.5316	**2.5845**
8	24.2938	8.3902	0.3257	1	0.5	1	1.0057	0.4595	**7.4577**
9	21.3578	9.055	0.4464	1	0.5	1	1.0248	0.4567	**6.3633**
10	22.064	8.7065	0.4023	1	0.5	1	1.0256	0.4397	**7.1653**

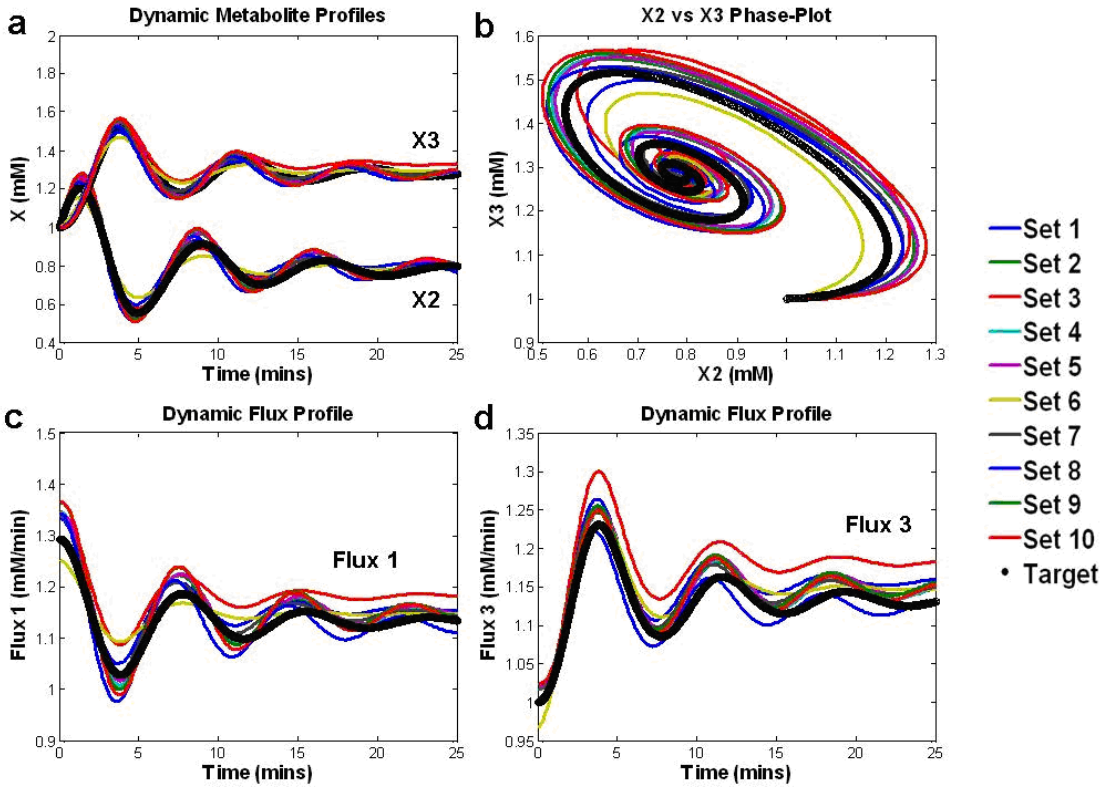


Figure 19: Variations in parameters may be compensated throughout the entire system and yield similar residual errors and data fits.

As an extension of error compensation within an equation, error may be compensated among different equations (Figure 19), thereby potentially spreading errors in some part of the model throughout the entire system. It is to be expected in this case that fluxes describing other variables are poorly represented.

Effects of error compensation on extrapolation

Suppose the same system is now studied with a different amount of input X_1. Specifically, assume that in a new set of experiments X_1 is reduced from 2 to 1.1, and the question is whether the estimated model is able to predict the dynamics of X_2 and X_3. Selecting parameter sets 1 and 6 from Table 4 confirms that both produce very similar responses for $X_1 = 2$. However, for $X_1 = 1.1$, the responses are quite a bit different (Figure 20).

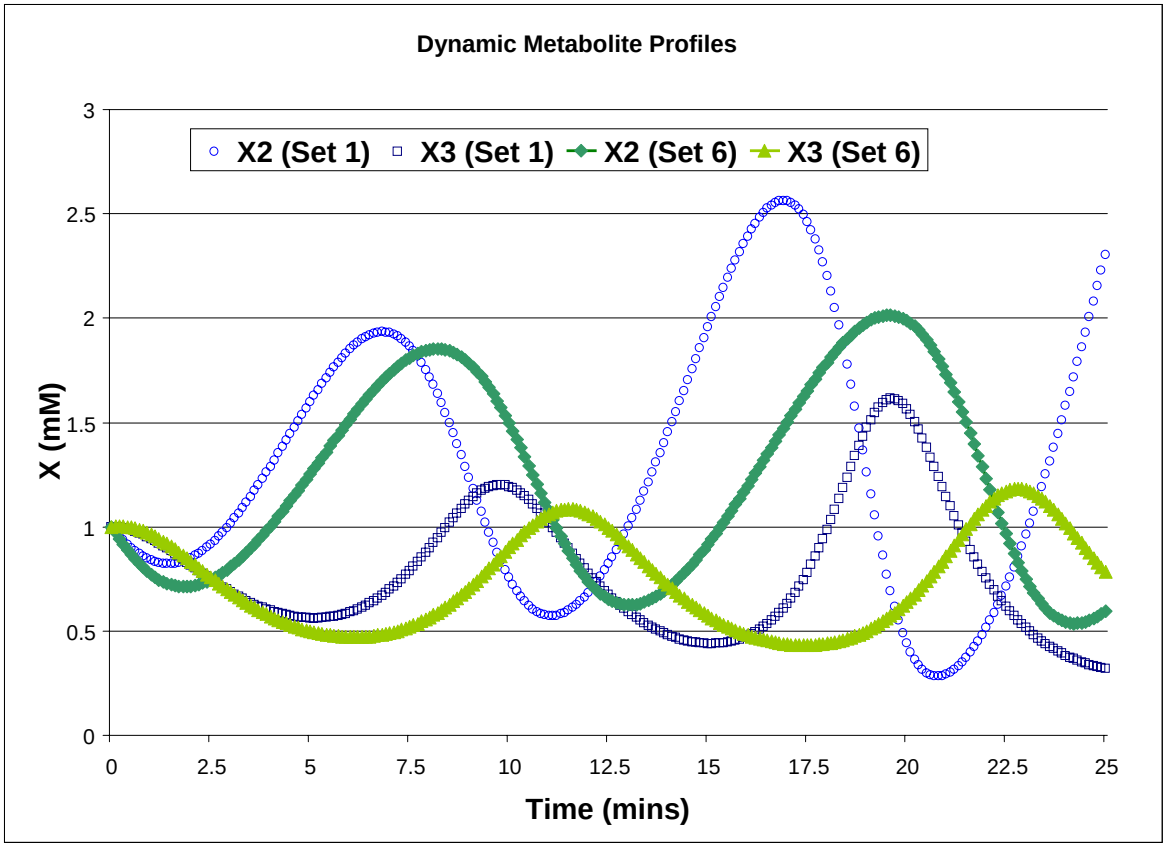


Figure 20: While two numerical system representations may be similar for the dataset used for fitting, they may lead to dynamic responses that are quite different if some of the "experimental settings" are changed. In the case shown, the input in pathway (figure 16) was reduced from 2 to 1.1.

These illustrations highlight some deep rooted flaws that have been inherent in parameter estimation strategies to date but as discussed ahead, these can now can be addressed with the innovative approach of DFE.

A novel approach

A novel approach to estimating metabolic pathway systems, called *Dynamic Flux Estimation (DFE)*, is proposed, which resolves several of the issues outlined and explained above. The approach consists of two distinct phases. The first consists of an entirely model-free and assumption-free data analysis that reveals inconsistencies within the data, and between data and the alleged system topology. The second phase addresses the mathematical formulation of the processes in the biological system. In contrast to all currently available methods, this phase allows quantitative diagnostics of whether—or to what degree—the assumed mathematical formulations are appropriate or in need of improvement. DFE builds upon the tenets of stoichiometric [81-83] and flux balance analysis (FBA; for a review see [84] in that it focuses on the stoichiometry at all nodes in the investigated system to ensure conservation of mass and to estimate flux distribution at each instant in time. However, in DFE the system is typically not in a steady state or quasi-steady-state [35-38, 85, 86], and its transient dynamics is utilized as a crucial indicator of the regulation within the system.

Because DFE consists of two phases that include several steps, some of which are new, some computational, some logistic (*e.g.*, the choice of mathematical representations in the second phase), and some using any of a variety of existing methods, its exact computational time requirements and accuracy of solution are difficult to assess against currently available methods. Nonetheless, results obtained from case studies (presented here) suggest that the proposed approach is more effective and robust than presently available methods for deriving metabolic models from time series data. Specifically, its combined model-free and model-based analyses avoid compensation among and within equations and therefore promise significantly improved extrapolability toward new data or experimental conditions. Its diagnostic tools pinpoint causes of inadequate fits between model and data and suggest either changes in assumptions related to model choice or the use of data as un-modeled "off-line data".

The following sections describe DFE and demonstrate its features with a series of successively more complicated (and more realistic) situations, beginning with an idealized, yet representative case, and ending with actual experimental observations describing fermentation in the bacterium *Lactococcus lactis*.

The proposed method requires time series data that characterize the dynamics of the system variables. Such data are still relatively rare but are being generated with increased frequency and quality. Some suitable data sets that exist already have been obtained with *in vivo* NMR [34], mass spectrometry [86] and other methods [87]. Furthermore, the prospect of the availability of efficacious methods of analysis may inspire experimentalists to generate more of these types of data, which is technically possible and probably worth the effort, even if it is more expensive. Since much of the advantage of DFE is the result of natural constraints among fluxes, DFE is particularly useful for metabolic systems, but less so for gene expression and protein interaction systems.

Method

DFE is a phased approach with well-defined outcomes for each step and rigorous checks and balances that ensure consistency of the solution (see Table 5 below).

Table 5: Phases and steps of dynamic-flux based parameter estimation from metabolic time series data

Phase	*Steps/Activities*	*Outcomes*	*Checks & Balances*
I	• Estimate missing data (if any) • Smooth and optimize data to achieve mass balance (if necessary) • Estimate slopes	• Smooth, balanced time-series data • Slope estimates	• Mass/Material balance
	• Formulate system of fluxes • Solve linear system of fluxes at each time point	• Linear system of fluxes • Dynamic flux profiles	• Flux balance • Integrateable flux time-series

II	• Evaluate flux-substrate plots to choose representative functional forms • Fit parameters of kinetic function to flux profiles	• Parameterized kinetic model	• Check fit of functional forms to flux profiles • Simulate system

Each phase facilitates incremental development and analysis of the metabolic target model. Phase I, which is entirely model-free, consists of two distinct sets of activities yielding slope estimates and dynamic flux profiles. First, the experimental data are analyzed for mass/material balance and smoothed as necessary. Slope estimates can be derived using different numerical techniques. Next, the pathway structure (*i.e.,* the system topology) is used to generate a system of symbolic equations describing the dynamics of the system. Substituting slope estimates in this system of equations results in a system of fluxes that is linear at each time step *t*. This linear set of equations can be solved at each time step to obtain dynamic (time-series) profiles of all fluxes in the system. These dynamic flux profiles can be checked for flux balances at the overall system level and at the level of each metabolite pool. Phase II is model-based. Here, based on the flux profiles from the previous phase, one evaluates each plot of a flux versus its alleged substrates and modulators to analyze and choose between possible mathematical representations for each flux. Once decided, the parameters of the chosen functional form are fitted easily with some regression technique to obtain a fully parameterized kinetic model for the system. The fitness of the parameters for each flux function can be evaluated independently and the same can be done for the overall system performance.

Wide arrays of robust numerical techniques are available for the computational aspects of each component of DFE, including data smoothing, slope estimation, the assessment of linear flux systems, and linear/nonlinear regression methods for parameter estimation. The proposed DFE workflow (Figure 21) consists of distinct steps.

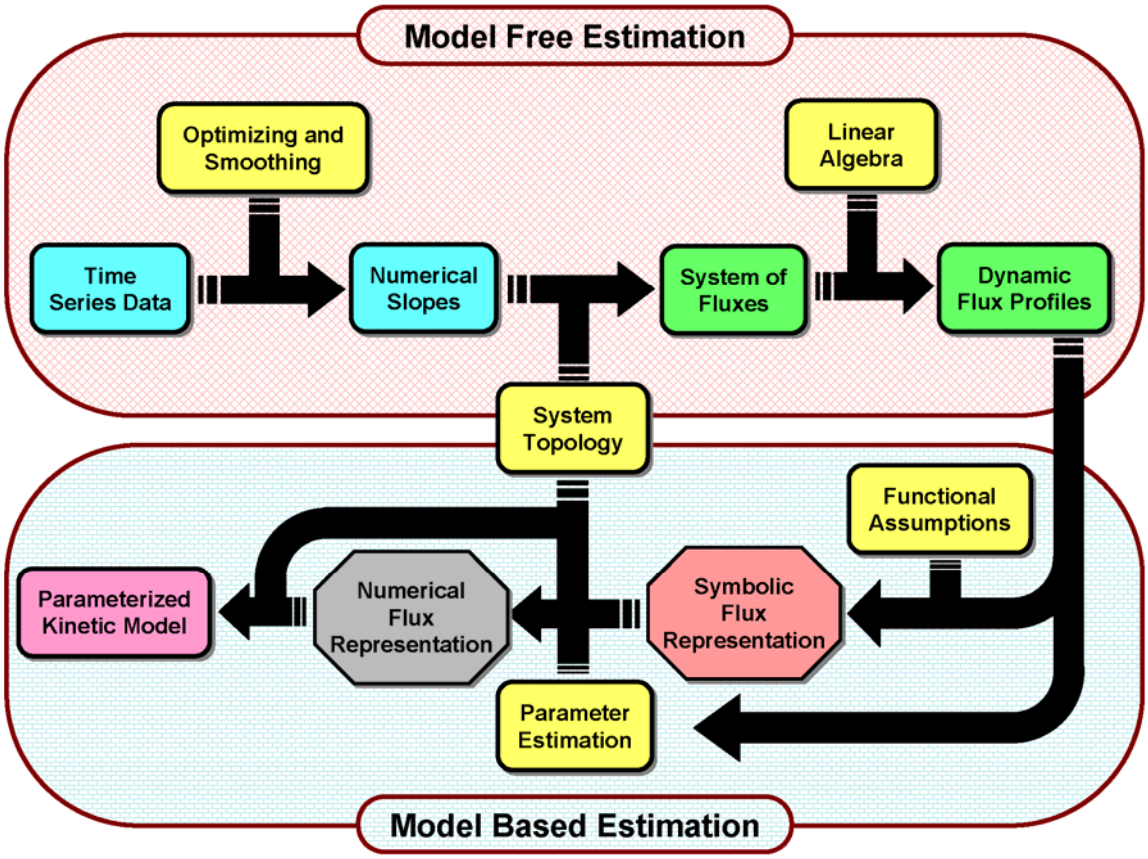


Figure 21: Dynamic-Flux Estimation (DFE) approach to metabolic system estimation from *in vivo* time-series data. Starting with experimental time series, the data are simultaneously balanced and smoothed for constant total mass throughout the time series. Then the slopes are estimated using published methods. Combined with the knowledge of the system topology, the slope information yields a linear system of fluxes. The system is solved, using linear algebra techniques, yielding dynamic profiles of all extra- and intra-cellular fluxes in the system. Next, functional assumptions are formulated on how to best represent the processes mathematically. These functions result in symbolic flux representations that can be independently fitted with regression methods to the respective dynamic flux profiles. When combined with knowledge of the system topology, the numerical flux functions are integrated as a single unified system model to obtain time courses.

1. ***(Phase I: Model Free Estimation)*** If necessary, smooth and balance the data in the sense that there should be no gain or loss of material over time. This balance is readily checked against the system stoichiometry. I developed for this purpose a combined non-linear programming and moving-average algorithm to remove noise while simultaneously balancing the time-series data for constant total mass. The smoothing and slope estimation aspects can be

accomplished with finite difference approximations, cubic splines, or more sophisticated methods [88].

2. ***(Phase I: Model Free Estimation)*** Substitute differentials with estimated slopes for each variable and at each time point [13, 30, 31] and construct a linear system of the form "Slope Vector(t) = [Stoichiometric Matrix]×[Flux Vector(t)]," where the matrix is directly derived from the known (or hypothesized) topology of the system. Solve the system with methods of linear algebra. The result is a (discrete) set of dynamic profiles (time series) of all extra- and intra-cellular fluxes in the system. Over-determined systems require the pooling of fluxes or the use of pseudo-inverse methods. Several constraint-based optimization techniques have been proposed for flux analysis of underdetermined metabolic networks [89]. These approaches have become a mainstay of FBA and served well under steady-state and quasi-steady state conditions [84, 90]. Analogous methods may be developed for DFE by using these established approaches as the starting point. Also, Ishii and collaborators recently proposed a hybrid method for modeling metabolic systems [91]. This novel approach distinguishes between dynamic and static enzyme activities based on the estimation of time dependent enzyme reaction rates. The system is split into dynamic and static modules such that a quasi-steady state is attained in the static module at each instant, while the complete system acts dynamically. The transient dynamics of the system is regenerated by interactions between kinetic-based dynamic models and metabolic flux analysis-based static models. A similar separation in dynamic and static modules could be applied to DFE as well. In addition, underdetermined systems may be complemented with information from steady-state FBA, concentration measurements using mass spectrometry or NMR, and traditional enzyme kinetics. Finally, it is possible to pool sequential and collinear

variables [50] and to combine DFE with methods of structure identification [33, 50] that are to be applied to select portions of the system.

3. ***(Phase II: Model Based Estimation)*** Up to this point no assumptions have been made with respect to the mathematical formulation of the flux terms. The next step is now to plot each flux against time and also against the variables affecting this flux (possible in two or three dimensions). As a default, assume that each flux V_k is representable as a product of power-law functions of form $R_k X_i^{f_{ki}} \ldots X_n^{f_{kn}}$ as it is done in Biochemical Systems Theory (BST [11, 13]). Regress V_k in logarithmic coordinates against the contributing variables to obtain the rate constant R_k and the kinetic orders $f_{ki}, \ldots f_{kn}$, etc. Analyze the quality of fit visually and/or with methods of linear regression diagnostics [92]. For non-power-law flux representations (*e.g.*, Michaelis-Menten or Hill functions), it might be possible to execute the analysis with inverse quantities, as in Lineweaver-Burk analysis, or one has to resort to methods of nonlinear regression.

Case studies

Application of DFE is demonstrated here with four case studies that were inspired by data describing how the bacterium *L. lactis* converts glucose into lactate via the pathway shown in Figure 22a. The data was obtained from *in vivo* Nuclear Magnetic Resonance (NMR) time series experiments with *L. lactis* that were performed by our collaborators Drs. Santos and Neves at ITQB, Portugal [34, 51, 93-98]. These NMR data, with a time resolution of 30 seconds, provided time courses for substrate consumption, product formation and intracellular metabolite pools, all monitored *in vivo*. Specifically, they characterized the dynamics of glycolytic metabolite pools in a suspension of cells that metabolized a 20 mM bolus of [6-^{13}C] glucose under aerobic conditions at pH 6.5 [34]. These data are as good as a modeler can presently hope for. They are more or less

complete, show clear trends and exhibit experimental noise that is quite reasonable in most cases. Most aspects of the observed time courses make intuitive sense. The bolus of external glucose is gradually used up, during which time all subsequent metabolites increase in concentration. With the external glucose pool becoming depleted, the immediately subsequent pools (G6P and FBP) decrease while the subsequent trioses (3PGA and PEP) approach high levels. Interestingly, these high levels do not decrease appreciably during the hour-long experiment, even though the pathway is essentially linear.

Idealized situation (proof-of-concept)

DFE was first tested and proven to work with an idealized data set (Figure 22b), which was constructed per simulation with an earlier model [61] (see Appendix A). These data are by design smooth and balanced and permit error-free estimation of slopes directly from the equations. Following the guidelines of DFE, the stoichiometric, time dependent matrix equation was solved using computed slopes on the left-hand side of this equation, and the flux values were thus obtained at each time point *t* (Figure 22c).

Note that these dynamic flux profiles were obtained purely from knowledge of the system topology and "experimental data," yet without any assumptions regarding an underlying functional model. Mimicking a realistic situation, a numerical model was derived based on the assumption that all fluxes could be validly modeled with products of power-law functions, as it is customary in BST. Thus using a symbolic power-law representation for each flux that included all contributing variables, the estimation of the kinetic orders and rate constant was straightforward since each flux term becomes linear when represented in logarithmic coordinates. The dynamic model with these flux representations was integrated and its behavior closely matched that of the experimental time-series data (Figure 22b). (see Appendix A for model details)

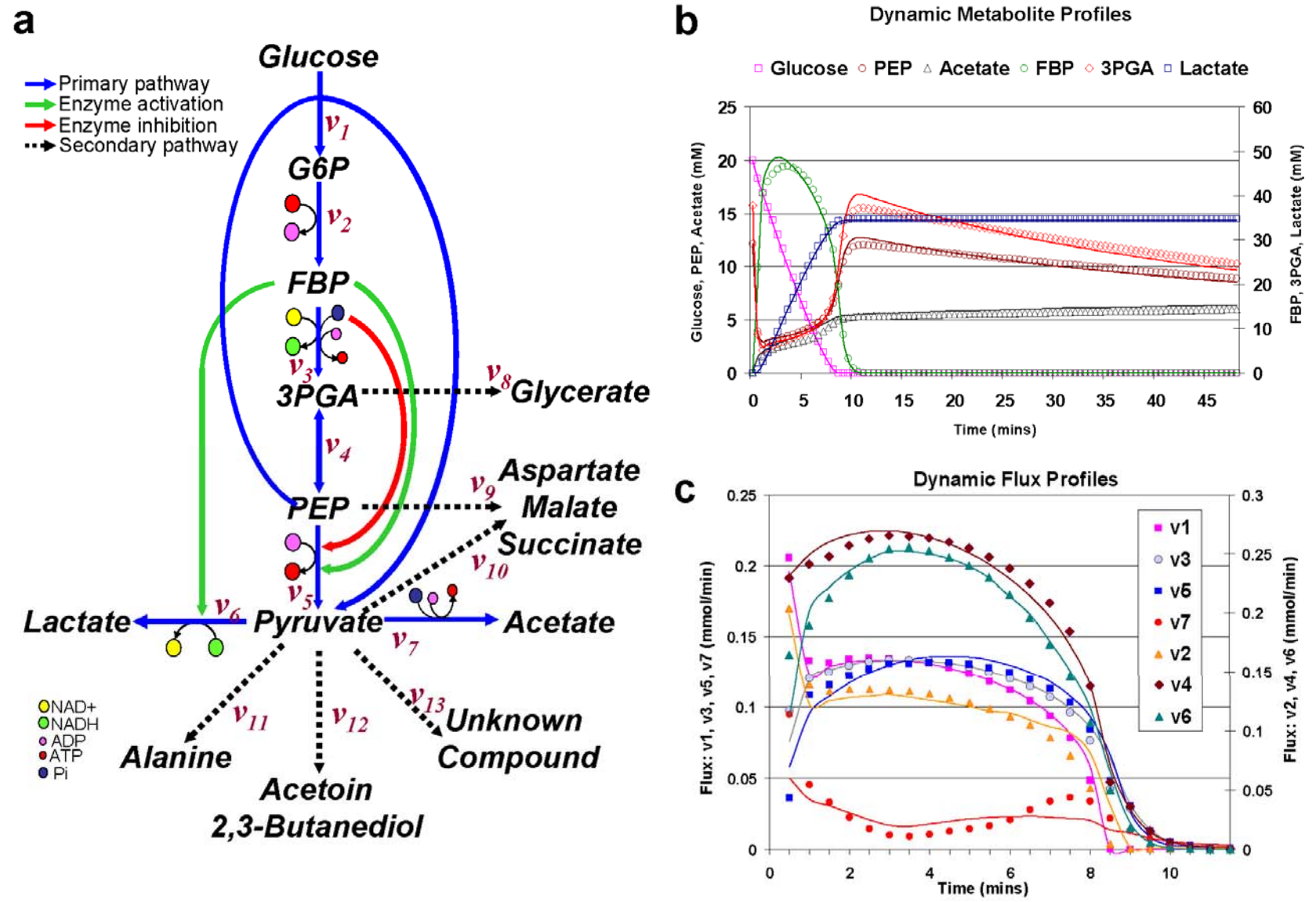


Figure 22: Results of Case Study 1. (a) Fermentation pathway in *Lactococcus lactis*. Dark arrows show flow of material. Dashed arrows indicate leakage of material into secondary pathways. Enzyme activation and inhibition are indicated by light gray arrows. Abbreviations: G6P: glucose 6-phosphate; FBP: fructose 1,6-bisphosphate; 3-PGA: 3-phosphoglycerate; PEP: phosphoenolpyruvate; ATP: adenosine triphosphate; ADP: adenosine diphosphate; Pi: inorganic phosphate; NAD^+: nicotinamide adenine dinucleotide (oxidized); NADH: nicotinamide adenine dinucleotide (reduced). (b) Dynamic metabolic profiles. Time series data of major metabolites in the primary pathway (symbols). Solid lines indicate fits with a model derived using DFE. (c) Dynamic flux profiles. The symbols show the time series of flux profiles estimated solely from data and the system stoichiometry using DFE. The solid lines indicate fitting of a power-law model to the dynamic flux data.

Simulated data with noise

To test the robustness of the DFE approach against noise, 10% artificial pseudo-random noise (drawn from a uniform distribution) was added to the ideal data set from Case 1. Due to the noise, the total mass in the system was no longer constant and required balancing, along with smoothing (Figures. 24a, b). An iterative optimization and smoothing scheme was developed to simultaneously smooth and balance the metabolic time-series data (Figure 24a). The slopes were then estimated from the smooth balanced data. Substituting slopes in the stoichiometric equation, the solution to the linear system of fluxes was obtained at each time point *t* (Figure 24c) and parameters were estimated for each of the power-law functional forms. The result was a fully parametric kinetic model (Figure 23) that captured the dynamic behavior of the noisy experimental data well (Figs. 24c, d).

Flux	Power-law Flux Models
v_1	$11.06(\text{Glucose})^{0.35}(\text{PEP})^{0.86}$
v_2	$2.28(\text{G6P})^{0.55}(\text{ATP})^{0.05}$
v_3	$1.33(\text{FBP})^{0.87}(\text{Pi})^{0.04}$
v_4	$21.46(\text{PEP})^{0.22} - 14.26(\text{3PGA})^{0.15}$
v_5	$26.64(\text{PEP})^{0.48}(\text{FBP})^{1.23}(\text{Pi})^{-0.001}+16.14(\text{PEP})^{2.47}$
v_6	$100(\text{Pyruvate})^{0.6}(\text{FBP})^{0.59}$
v_7	$500(\text{Pyruvate})^{0.86}(\text{Pi})^{0.73}$

Figure 23: Numerical power-law model derived for the scenario in case study 2 (simulated date with noise)

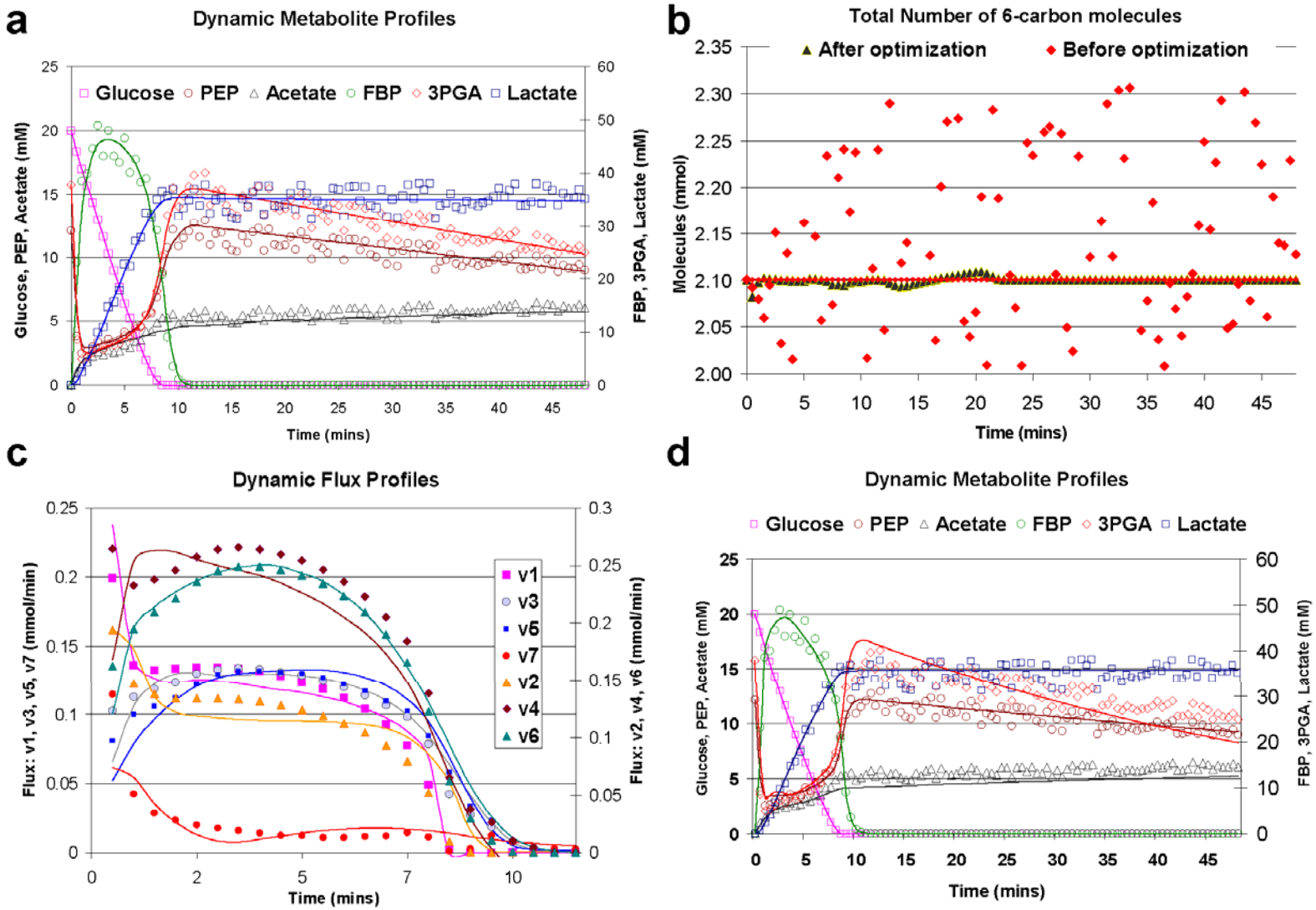


Figure 24: Results of Case Study 2. (a) Dynamic metabolic profiles. Metabolic time-series data with added artificial noise (symbols). The solid lines represent the smoothed and balanced time series. (b) Dynamic mass balance. The random noise leads to mass imbalance which is successfully restored after optimization and smoothing. (c) Dynamic flux profiles. The linear system of fluxes is solved to obtain unique flux profiles (symbols). Power-law models are fitted to each flux time series independently (solid lines). (d) Results from the numerical model. Using DFE, a fully parametric kinetic model is derived from noisy metabolic time series data (symbols). The results of the model (solid lines) closely match the original dynamic metabolic data.

Simulated data with non-power law terms

In the first two cases, the data generating system was implemented with power-law representations. To test and demonstrate the diagnostic capabilities of DFE, the same system (without noise from case 1) was simulated with a non-power-law, sigmoidal glucose uptake function (Figs. 25a, b). Next, the slopes were estimated and the dynamic stoichiometric system was solved as before. The estimated fluxes were notably different from those obtained in the earlier studies, especially at the initial time points (Figure 26). Attempts to model this system of fluxes exclusively with power-law functions failed. Other methods would have had to stop at this point, simply concluding that the fit was sub-optimal. Even worse in some sense, the simultaneous fitting of all equations or of all terms within each equation would have led to error compensation between terms, thereby not only mis-fitting the sigmoidal flux but other fluxes as well. The overall fit might actually have been acceptable, but attempts to extrapolate the resulting numerical model to other datasets or conditions would have become problematic. In contrast to this "system-wide distribution of error," DFE prevented such distribution of error and pinpointed the source of error accurately by enabling me to test every flux individually against any hypothesized functional representations. When executing this analysis with power laws, using linear regression in log space, the result was very encouraging: All fluxes were reasonably well represented with power-laws except for the uptake process (Flux v1). Evaluation of the flux plots for this reaction step (Figure 25c) confirmed that the flux in glucose and PEP deviated systematically from the experimental flux when it was modeled by a product of power-law functions. More importantly, even though this flux was not well represented by power-laws, I obtained excellent power-law fits for the other fluxes, such as flux v3 (Figure 25d), which clearly demonstrated that errors in one flux were not compensated anywhere else in the system.

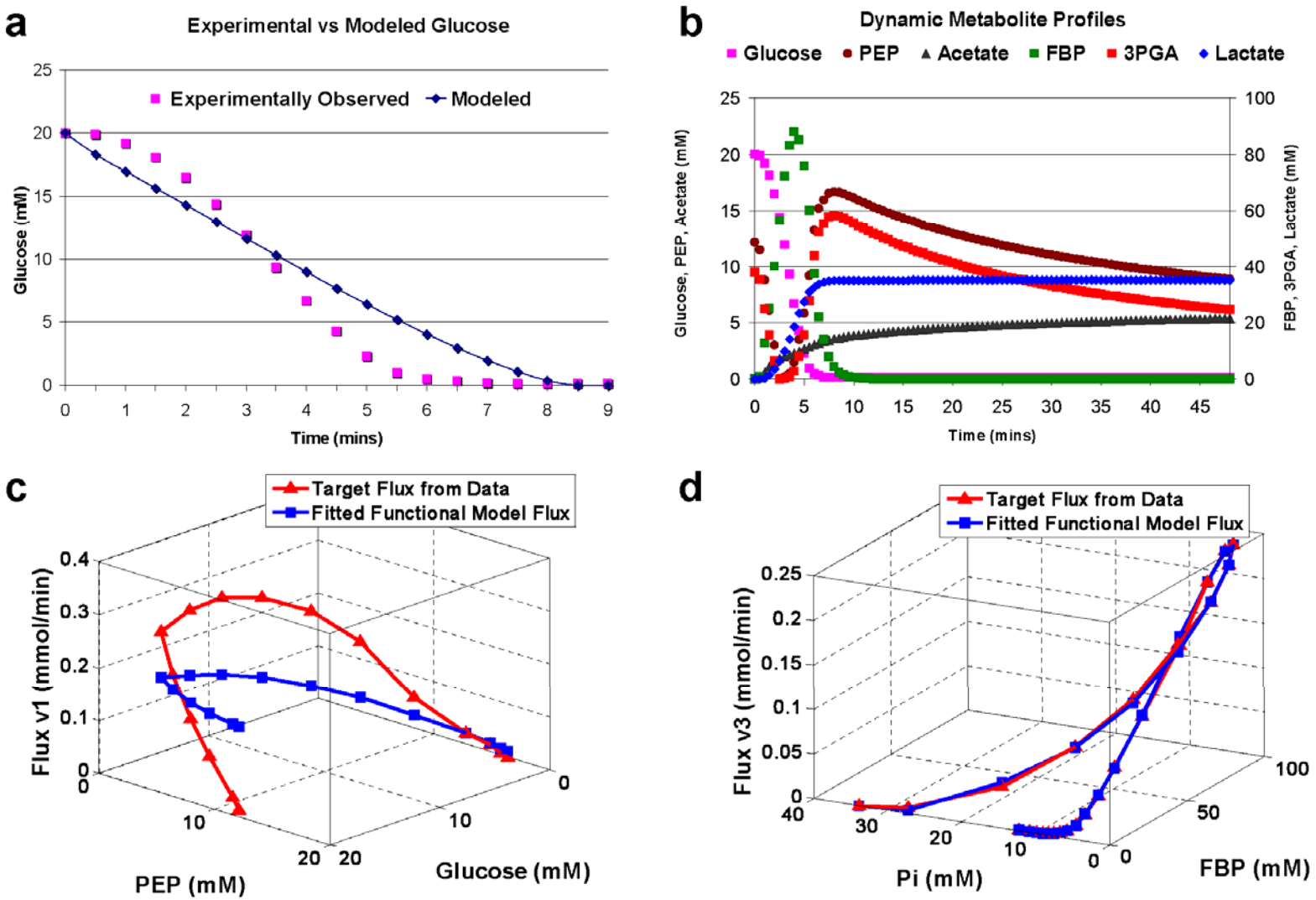


Figure 25: Results of Case Study 3. (a) Sigmoidal glucose uptake. This type of uptake dynamics has been observed in experiments (symbols) and is difficult to represent with a simple power-law function (solid line). (b) Dynamic metabolic profiles. Time series data of the major metabolites that result from sigmoidal glucose uptake. (c) and (d) Flux substrate plots. The "experimental" flux profile (red), obtained using DFE, is plotted against the corresponding flux obtained by fitting a power-law model (blue). Fig. (c) shows systematic error when flux v1 is fitted with a power-law model. On the other hand, a power-law model accurately reproduces other fluxes like v3 in the same system (d).

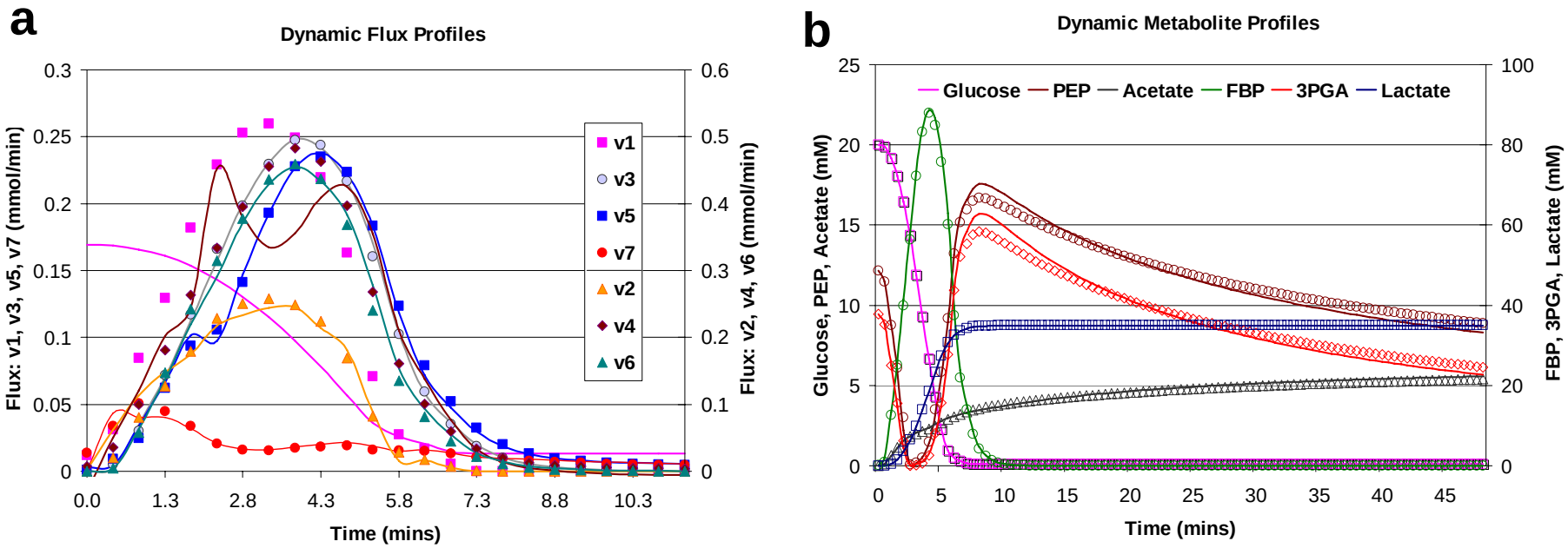


Figure 26: (a) Dynamic flux profiles estimated purely from data (symbols) and flux profiles, modeled as power-laws (solid lines) for Case scenario 3. It is obvious that flux v_1 (glucose uptake) is not well modeled. (b) Dynamic metabolite profiles (symbols) simulated with the sigmoidal sugar uptake model and results from a model, derived using DFE, where glucose uptake was flagged as non-power-law and therefore taken off-line; all other fluxes were represented as power-laws (solid lines).

Real data

Many methods seem to function well for artificial data, yet break down in the real world. Therefore actual experimental NMR data from the *L. lactis* pathway (Figure 22a; Figure 27a) was used to further test DFE. Note that experimental measurements were available for only key metabolites (including G6P, FBP, 3PGA and PEP). The other metabolites along the primary pathway, such as F6P, DHAP, GAP etc could not be measured as they were well below the detection limit of the NMR. Also, data for the initial five minutes, for both 3-PGA and PEP, were made-up i.e. they are artificial data points. In reality, PEP and 3PGA are not detected before addition of labeled glucose, because they are unlabeled, but after the glucose bolus and while glucose is present they are not detected because their concentration is below the detection limit. Strategies to deal with the peculiarities of this data set have been discussed in detail in [61].

As a first check, the total mass in the raw experimental data was assessed at each time point and it was detected that they were significantly unbalanced (Figure 27b). None of the current parameter or system estimation algorithms, including our own [33, 61, 88], check for overall mass balance. As a consequence, these algorithms model something different from what is implicitly expected, which casts doubt on the ultimate estimation results and is likely to lead to problems with new data sets or extrapolations.

Attributing the imbalance to measurement noise merely did not allow for mass balancing within acceptable noise limits. Consultation with our collaborators revealed that several secondary metabolites and fluxes had not been included in the main dataset Figure 22a). These metabolites are not as well characterized as the mainstream species because they are present in very low abundance and hence unobservable during the course of the experiment. Using single time-point measurements, obtained at the end of the experiment, and combining knowledge of the pathway for these minor metabolites, the expected time profiles of these metabolite concentrations were estimated (Figure

28a). Accounting for these minor metabolites finally enabled me to balance the system in mass (Figure 27b).

Once mass balanced data was achieved, subsequent steps of DFE were executed to compute slopes, estimate flux values at each time point *t* (Figure 27c), and fit all fluxes, except for glucose uptake, by power-law functions. Instead of trying to fit the uptake with some sigmoidal function, this flux was left un-modeled and incorporated it into the model as off-line data [57, 61]. All other fluxes were found to fit very well (Figure 27c) but yet when integrated the system did not yield the same results as observed in experiments. Upon close inspection it was found that minor deviations in the metabolic profiles during integration caused the flux functional forms to return values which were far from the true (DFE) fluxes. Because these true fluxes were known, I replaced each flux function by an offline spline of that flux (one at a time) and integrated the system back again. By using this somewhat reductionist approach, the source of error was located to be present in the functional form for flux v2 (Figure 28c) which was causing the G6P and FBP profiles to be erratic during simulation. Additionally, when trying to fit a functional form for the flux between PEP and Pyruvate, it became apparent that the conversion of PEP into pyruvate may consist of two fluxes (in addition to the PTS system), namely the main flux that is subject to activation by FBP and inhibition by Pi and a very small flux that is less affected by these modulators. The reason to postulate this minor flux is the observation that acetate continues to increase even after FBP is depleted. To model this situation, the flux v_5 therefore consists of two components. In the end, the result was a parametric kinetic model that closely reproduced the dynamics of the metabolite pools (Figure 27d). It is worth noting that the residual error of this model may be larger than the error in a model that is optimized with standard methods, because a standard estimator has the freedom of distributing errors throughout some or all fluxes, which DFE does not permit. As a consequence, the total error in DFE may be higher, but the fit to each individual flux is more reliable.

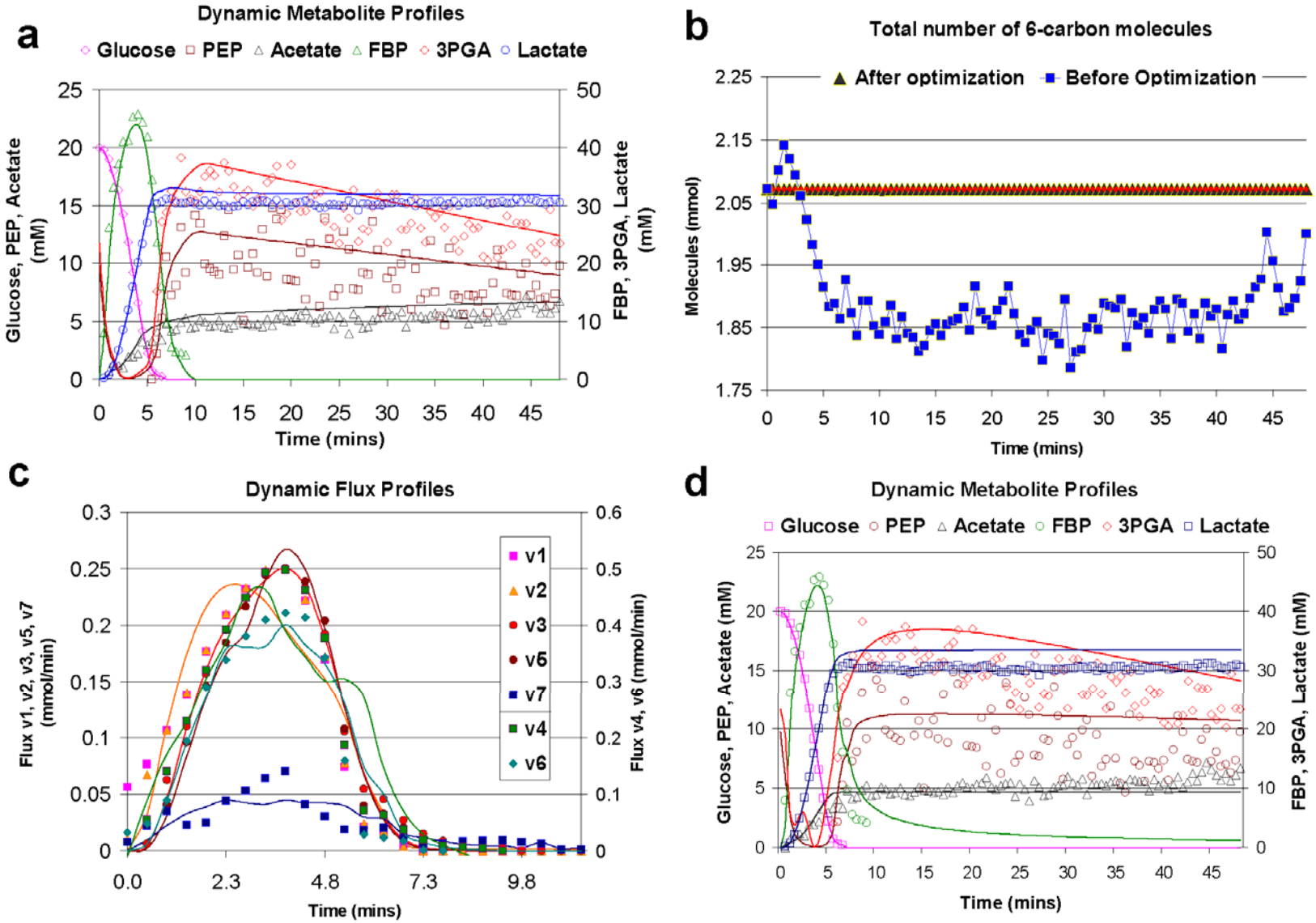


Figure 27: Results of Case Study 4. (a) Dynamic metabolic profiles. Measured dynamics of metabolite pools in *L. lactis* following a 20 mM [6-13C] glucose bolus (symbols). (b) Dynamic mass balance. Systematic mass imbalance in the experimental data was attributable to missing information about secondary metabolites. The balance was successfully restored by accounting for secondary fluxes. (c) Dynamic flux profiles. The linear system of fluxes is solved to obtain the unique flux profiles (symbols). Power-law models are independently fitted to each flux time series, using linear and non-linear regression (solid lines). (d) Results from the numerical model. Using DFE, a fully parametric kinetic model is derived from the actual metabolic time series data (symbols). The results of the model (solid lines) closely match the data.

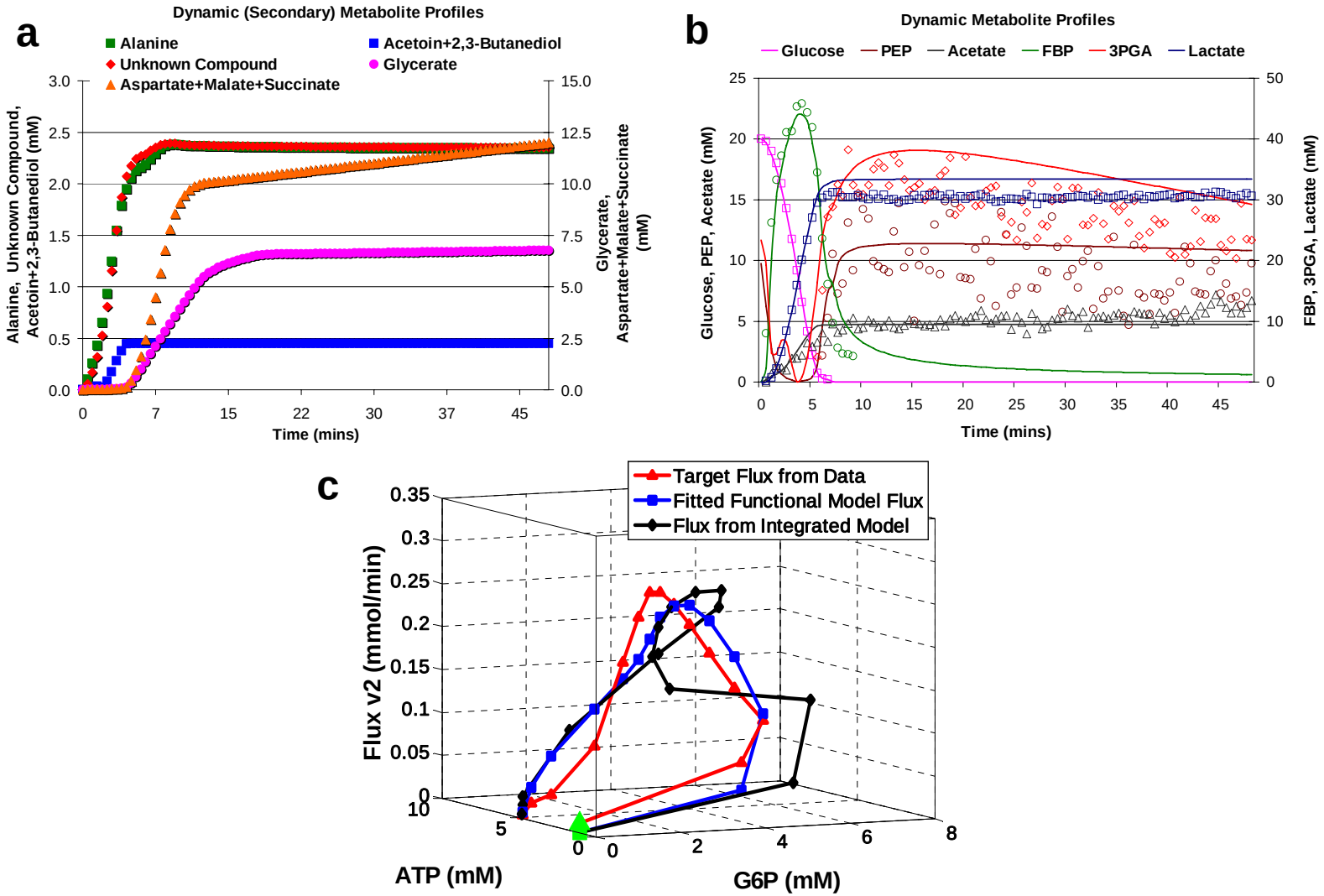


Figure 28: (a) Expected dynamics of secondary metabolites. (b) Actual data and model fit obtained upon mass balancing, consideration of secondary metabolites and application of DFE. (c) Pseudo-three-dimensional representation of flux v_2. Interestingly, the flux fitted to the data (red) in the model-free phase of DFE is quite close (blue). Although the flux in the fully integrated model is less well modeled (black), the overall dynamics of the system (panel b) is acceptable. The green points indicate initial conditions.

Discussion

Dynamic Flux Estimation is proposed as a new approach that resolves at least some of the open issues in the estimation of metabolic pathway systems. The first, model-free and essentially assumption-free phase of DFE permits consistency checks within the metabolic time series data and leads to numerical representations of fluxes as functions of the variables affecting them. The second, model-based phase allows the objective testing of functional forms for fluxes and is not within the repertoire of any of the existing methods. The two-phased approach thus permits rigorous, quantitative diagnoses of the metabolic data, the alleged pathway structure, the assumptions made in the choice of flux representations, and the causes of residual errors. DFE eliminates compensation of error among terms and among variables, which has been a tremendously complex problem with other methods, especially when it comes to extrapolations with the estimated model.

While DFE very significantly reduces error compensation between equations and between flux terms, it still admits error compensation among the parameters within a given flux, independent of what representation is chosen. In the context of BST, this type of compensation between a rate constant and the kinetic orders is well known [33, 80, 99]. For reliable extrapolations, the within-flux compensation should also be removed. This removal seems to require data covering wide ranges of variation, multiple datasets or additional information about some of the parameter values, for instance, from traditional enzyme kinetics.

It has been observed in related work that the strategy of replacing differentials with slopes may lead to good fits for the dynamics of each variable in isolation, yet cause problems when all estimated parameter values are entered into the differential equation model [30]. The reason is that even small deviations between data and model results in one variable can lead to an amplification of error in other equations. This issue occurs in DFE as well. However, in contrast to other methods, DFE allows diagnostic analyses of

the solution (Case of flux v2 in scenario 4; Figure 28c). In response to such a situation, one may ignore the differences, search for causes of the deviations, or substitute smoothed data for a troublesome flux in the form of an off-line process [57, 61].

A key feature of DFE is the requirement of time series data that are sufficient to capture the dynamics of the system. It is in general difficult to say how many data points are needed for reliable estimations. The key reason is that there is no good, quantitative criterion for the complexity of a time course. In simple dynamic responses, such as monotonically saturating functions, a few data points may be enough to characterize a time trend with sufficient reliability. In other cases, such as the example demonstrated here, the number of time points needed is higher. It seems quite evident that the number very much depends on the complexity of the time course and the noise in the data. Importantly, the types of data required for DFE are becoming more commonplace because modern methods of molecular biology permit their measurement with a variety of already existing experimental methods.

DFE is an estimation approach particularly geared towards metabolic pathway systems, which are better suited for this type of estimation than genomic or proteomic systems because of conservation of mass at all nodes. Furthermore, DFE focuses on parameter estimation rather than on the identification of structure and regulation in ill-characterized pathway systems. Issues needing further development are related to missing data, missing flux information, underdetermined stoichiometric matrices, and ill-characterized systems topologies.

CHAPTER 4

COMBINING MULTIPLE DATA SOURCES WITH DFE [3]

Under ideal conditions, DFE appears to be as close to perfect as it is currently possible. However, it has two very significant limitations: i) DFE requires comprehensive time-series data, which are seldom available, and ii) the linear system of fluxes needs to have full rank. This chapter discusses how these issues may be overcome by resorting to information from additional sources.

Complementation of DFE with additional information

A direct, unique solution of the flux equations in DFE is only possible if the flux system is of full rank. The most frequent case in practical applications, however, is an under-determined system, because most actual pathway systems contain more fluxes than metabolites. As a consequence, the best purely algebraic solution possible is the expression of some fluxes as functions of other fluxes, which is not very useful *per se*. However, in most practical cases, other information about the system is known, and this information may be used to complement DFE. This complementation does not come for free and either requires assumptions about functional forms of fluxes, mechanistic details, or inferences regarding missing time series.

As an illustration, the glycolytic pathway in the bacterium *Lactococcus lactis* (Figure 29) is used. The experimental data (Figure 30) were obtained from the laboratory of Drs. Helena Santos and Ana Rute Neves who utilized the method of *in vivo* nuclear magnetic resonance (NMR) to measure the accumulation of intracellular metabolites

---

[3] Part of this chapter is published in: E. O. VOIT, G. GOEL, I.-C. CHOU and L. L. FONSECA, "Estimation of metabolic pathway systems from different data sources". IET Syst Biol. (In press), 2009

under anaerobic conditions following a 40 mM glucose bolus [34, 93]. Since glucose 6-phosphate (G6P) was not measured for this specific bolus, it was adapted from a corresponding NMR experiment with a 20 mM glucose bolus (presented in earlier chapters). Thus, data on the key metabolites (including G6P, FBP, 3-PGA, and PEP) were available, but data on less important metabolites (such as F6P, DHAP, GAP etc.) were not.

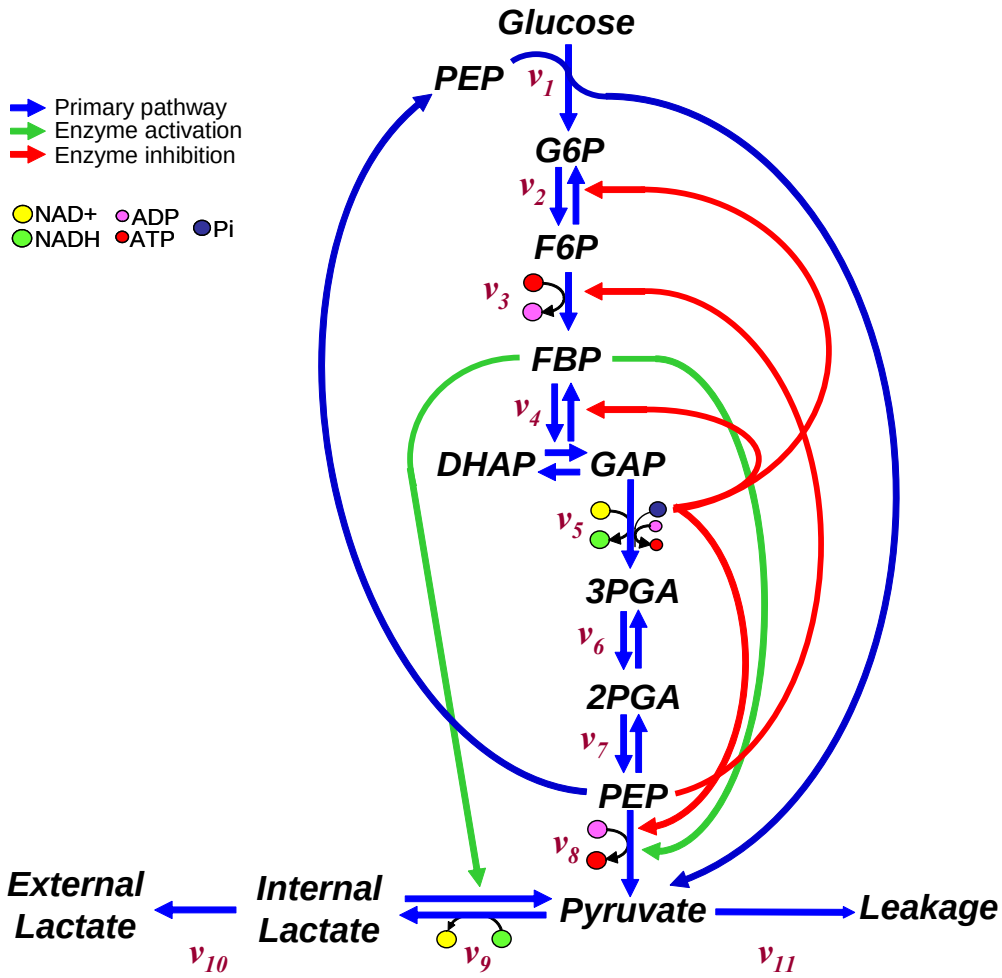


Figure 29: Schematic representation of the glycolytic pathway in *Lactococcus lactis*. Blue arrows indicate material flux, green arrows activation, and red arrows inhibition. Abbreviations: G6P: glucose 6-phosphate; F6P: fructose-6-phoshpate; FBP: fructose 1,6-bisphosphate; DHAP: dihydroxyacetone phosphate; GAP glyceraldehyde 3-phosphate; 3-PGA: 3-phosphoglycerate; 2-PGA: 2-phosphoglycerate; PEP: phosphoenolpyruvate; ATP: adenosine triphosphate; ADP: adenosine diphosphate; Pi: inorganic phosphate; NAD^+: nicotinamide adenine dinucleotide (oxidized); NADH: nicotinamide adenine dinucleotide (reduced).

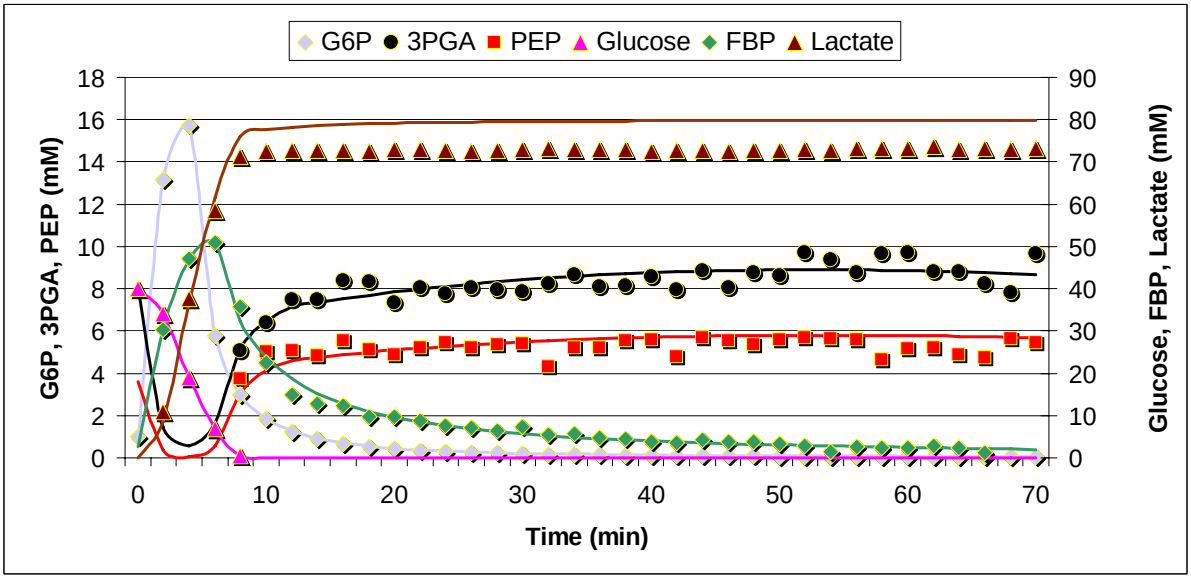


Figure 30: *In vivo* NMR measurements of metabolites of the glycolytic pathway in *Lactococcus lactis*. The symbols represent the raw experimental data. The lines indicate the output obtained by numerically integrating the system of DFE fluxes.

In generic terms, non-ideal situations that require complementation of DFE arise from a combination of the following issues.

Issue 1: The connectivity of the system is not fully known.

Issue 2: Some time series were not measured, although it is known that the corresponding metabolites are involved in the pathway. A typical example for this situation is a metabolite that is very quickly converted into another product, thereby precluding accurate measurements.

Issue 3: Some unknown or not measured metabolites are in fact important. The exclusion of these metabolites is a potential reason for mass imbalances in the system.

Issue 4: All relevant metabolites have been measured as time series, but the flux system is under-determined. This situation is the rule rather than the exception.

Resolving these issues seems only possible if additional information is available and/or if assumptions are made regarding the functional forms of some of the fluxes in the system.

Solution strategies for issue 1

Distinctly different methods have been developed for computationally inferring the unknown or ill-characterized connectivity of biological pathways (for a recent review see [100]). They include a wide spectrum of techniques, ranging from causality models [101-103] to perturbation methods [104], correlation based approaches [105], and probabilistic graph models for deducing causality [106]. Some methods (e.g., [26, 53, 56, 76, 107-109]) used time series data as the basis for their analysis. Specifically for metabolic pathways, methods like Alternating Regression (AR) [33] and Eigenvector Optimization (EO) [50] were proposed as structure identification methods that do not necessarily require knowledge of the connectivity or regulation of the pathway system.

If information is scarce or if the data are noisy, purely computational estimations are not always reliable, and within-term, within-equation, and between-equation error compensation may become a significant issue (illustrated in previous chapter). Instead of relying on structure identification algorithms alone, it may be useful to employ simpler algorithms that merely attempt to establish the connectivity pattern within the pathway. An example is a linearization procedure that generates probabilities for a given equation to be affected by combinations of system variables [26]. A different approach consists of an algorithm that reconstructs equations from the bottom up, testing first the data fit with the most parsimonious parameter set and gradually increasing the complexity of the equation [56]. It is also possible to optimize parameters for a predefined set of biochemically feasible candidate models [109].

Solution strategies for issue 2

The lack of time series data for certain metabolites may or may not be serious. An important determinant is the mass of the missing metabolite pools during the experiment. If this mass is small, methods of compensatory mass balancing may provide a solution that is not overly damaging. However, significant amounts of missing mass cannot be

ignored. If enzymatic information is available for fluxes producing and degrading a metabolite in question, it is sometimes possible to reconstruct its unknown time profile from neighboring time series.

Example: Consider the reversible isomerization of G6P to F6P, which is catalyzed by phosphoglucose isomerase (PGI). The kinetics of PGI has been characterized for both directions, and if one assumes a reversible Michaelis-Menten rate law for the net flux (Eq. 6), pertinent parameters are readily obtained from the literature [110-112]. By combining this kinetic *in vitro* information with the time series data on G6P and the *in vivo* G6P degradation flux estimates for v_2 at the measured time points, which was obtained with DFE, one can deduce the time series for the unknown metabolite F6P. This is accomplished by expressing Eq.6 with F6P as the dependent variable and solving it for all measured time points.

$$v_2 = \frac{v_{\max}^{for} \cdot \frac{[G6P]}{K_{mG6P}} - v_{\max}^{rev} \cdot \frac{[F6P]}{K_{mF6P}}}{1 + \frac{[G6P]}{K_{mG6P}} + \frac{[F6P]}{K_{mF6P}} + \frac{[P_i]}{K_{mP_i}}} \quad \text{..... Eq. 6}$$

The reconstructed F6P profile is similar to the G6P profile (Figure 31), but at a scale of about 1:10, which is in line with the common understanding of a fast equilibrium between the two.

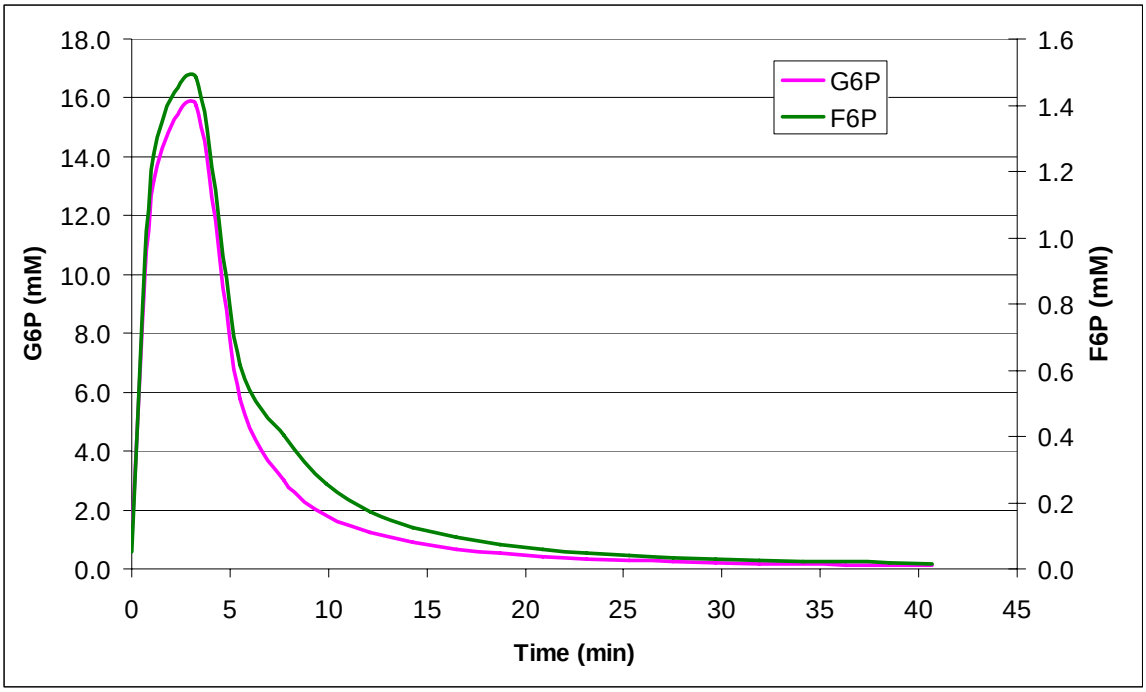


Figure 31: *In vivo* NMR measurements of G6P in *Lactococcus lactis* and reconstructed time series of F6P derived from a combination of DFE and kinetic literature information

Solution strategies for issue 3

The consequences of unknown or not measured metabolite pools may range from irrelevant to utterly detrimental for any estimation effort, depending on the extent of lacking information. A diagnostic aid for this situation is the checking of mass balance in the entire system throughout the experimental time period. If significant changes in balance are observed, because non-negligible amounts of mass are gained or lost, additional biological insight will be needed to remedy the situation. If the masses are more or less balanced, it is still possible that important fluxes or metabolites are missing. There is currently no obvious defense in this situation.

A slightly different situation occurs if relevant cofactors or modulators were not measured. For instance, NAD^+ and NADH may affect the speed of a reaction, but because of moiety conservation, no change in (carbon) mass is observable, so that the (carbon) mass in the system is perfectly balanced. Nonetheless, factors influencing the

NAD^+ / NADH ratio may significantly affect the dynamics of the pathway. Again, this situation requires a case-by-case treatment.

Example: As discussed in previous chapter (test case 4), the detected mass imbalance was too severe to be attributable to acceptable measurement noise, and smoothing efforts still left 10% of the supplied glucose unaccounted. It turned out that several secondary metabolites and fluxes had not been included, and accounting for these enabled the balancing of the system.

Solution strategies for issue 4

If the flux system is under-determined, it is necessary to obtain some fluxes by means outside DFE. Distinct options are available for this purpose, at least in principle. First, it may be possible to obtain fluxes directly from experiments. In a few cases, flux-substrate relationships were measured (*e.g.*, see parameter estimation in [48] from flux data in [113]), but such data are rare. Much more prevalent is information on the kinetic properties of enzymes and the reactions they catalyze. This information is closely linked to an alleged functional form for each flux. For instance, if a Michaelis-Menten rate function is deemed appropriate and if applicable K_M and V_{max} values can be found, the parameters and the time series data may be entered into the rate function to compute the appropriate flux value at each time point.

As an alternative, or if pertinent kinetic information is unavailable, it has been shown that regression methods, genetic algorithms (GA), as well as specialized methods like AR and EO [33, 50], have the potential of determining parameter values in pathway models from metabolic time series data [100]. This feature renders it possible in principle to determine the necessary number of missing fluxes and to use them in the first phase of DFE. A considerable drawback of this strategy is that GA and the various regression methods must *a priori* assume specific mathematical representations of the fluxes that are to be estimated. However, the most appropriate representations are often unknown. This

situation becomes less of a hindrance if some of the variables and fluxes operate within relatively small ranges, because one might expect that the typical canonical approximations, such as products of power-law functions or lin-log expressions, would be sufficiently accurate throughout these limited ranges. Thus, while many combinations of fluxes could theoretically be chosen to supplement DFE in an under-determined estimation task, it is advisable to choose variables and fluxes that remain relatively close to some normal operating values. At the same time, variables that do not vary much at all (i.e. that they stay at more or less a steady state) contain relatively weak information, which may lead to mis-estimation, so that the choice of fluxes requires a compromise. In addition to the fact that estimation algorithms must assume specific functions, they are also susceptible to error compensation between the terms of an equation.

Example: As an illustration for the use of kinetic information, pretend that the glycolytic system under investigation were under-determined. Specifically, consider the phosphofructokinase (PFK) step (v_3 in Figure 29), in which a phosphoryl group is transferred from ATP to F6P, yielding FBP and ADP. Since F6P is not observed under the given experimental conditions, it is not possible to estimate the PFK flux directly from the given time series data using DFE. However, it is well established that G6P and F6P are in rapid equilibrium, and because F6P is below the detection limit (2.5mM), it was assumed that its accumulation pattern is one-tenth that of G6P at all time points. This is a safe assumption considering the results that we have seen from the example demonstrated for issue 2 in the previous section (see Figure 31). It is furthermore known that the PFK reaction is essentially irreversible under physiological conditions and that the enzyme is allosterically inhibited by ATP, FBP and PEP, while being activated by ADP. Several rate laws have been proposed for the PFK reaction (*e.g.*, [114, 115] and references therein). I chose the model of Hoefnagel and collaborators (Eq. 7; [111]), because it was developed specifically for *L. lactis* under comparable conditions.

$$v_3 = \frac{v_{\max}\left(1-\frac{PEP^{n3PEP}}{PEP^{n3PEP}+K_{mPEP}{}^{n3PEP}}\right)\left(\frac{[F6P]}{K_{mF6P}}\right)^{n3}\left(\frac{[ATP]}{K_{mATP}}\right)}{\left(1+\left(\frac{[F6P]}{K_{mF6P}}\right)^{n3}+\frac{[FBP]}{K_{mFBP}}\right)\left(1+\frac{[ATP]}{K_{mATP}}+\frac{[ADP]}{K_{mADP}}\right)} \quad \text{..... Eq. 7}$$

Using this model with the published parameter values [111] and the time series of G6P (divided by 10 for the expected profile of F6P), a parameterized, mechanistic PFK model is obtained that very well represents the process *in vivo*, as it was obtained with DFE (Figure 32). This result is quite remarkable, first, because it confirms that kinetic information can indeed be used under opportune conditions to supplement DFE and, second, because it confirms that the entirely model-free phase of DFE yields very reasonable, numerical flux representations.

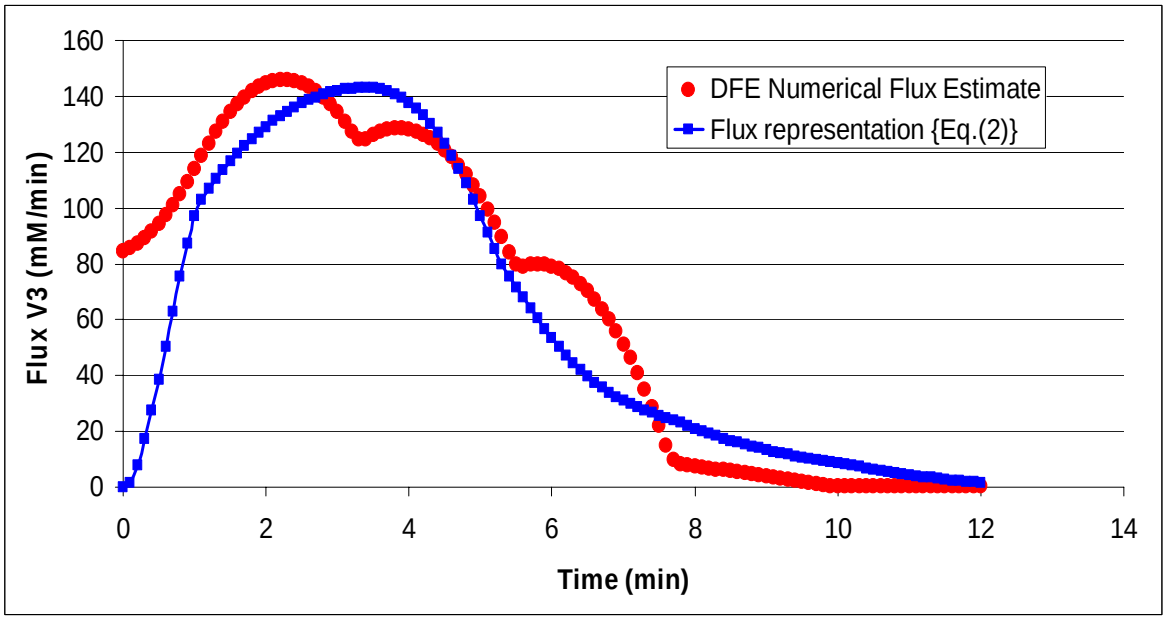


Figure 32: Flux v_3, obtained with DFE as a numerical estimate (red), and formulated as a published rate function with parameter values directly taken from the literature (Eq. (2); blue line). The numerical DFE estimate reflects different phases of glucose uptake, which may be due to a differential affinity of the cellular transporters to the α and β forms of glucose.

Discussion

DFE has the severe limitation that the fluxes in the pathway have to form a system of full rank. For more or less linear pathways, this assumption may be true, but as soon as pathway systems with cycles are under investigation, DFE cannot be applied directly, because the fluxes outnumber the metabolites. Other complicating factors are missing time series and uncertainties with regard to structure and regulation of the metabolic pathway system. Faced with the situation that ideal scenarios allowing direct application of DFE are rare, this chapter explored the question as to what degree DFE may be supplemented with other information. In essence, the ideas and solutions presented here suggest using DFE as a means to an integrative "bottom-up" and "top-down" system identification approach. The options for DFE supplementation span a range of methods. If all significant metabolic time series are available, and if some of the enzymes in the system are well characterized under pertinent conditions, it may be possible to construct flux-time and flux-variable relationships and use these as substitutes for unknown fluxes in DFE. Sufficient kinetic information may even allow the construction of time series profiles of metabolites that were not measured. Alternatively, or in addition, if one may reasonably assume functional forms for a few of the fluxes within the system, then a genetic algorithm or more specialized methods like Alternating Regression or Eigenvector Optimization can be employed to estimate a sufficiently large subset of fluxes to execute DFE on the rest of the flux system. The combination of methods presented here serves primarily as a proof of concept, and it is to be expected that targeted work on combined forward and inverse estimation methods will lead to refined and possibly even entirely novel system identification strategies. Such strategies will become increasingly important, because one should expect a rapidly growing number of time series data of high quality, which however will very seldom be comprehensive enough for a unidirectional estimation approach.

CHAPTER 5

A KINETIC MODEL OF GLUCOSE METABOLISM IN *LACTOCOCCUS LACTIS*

The integrative top-down and bottom-up approach to system identification, presented in the previous chapter, is applied to develop a detailed kinetic model of glycolysis. This case study presents a significant challenge for flux and parameter estimation since it has several elements of a real-life scenario: noisy and incomplete data, unobservable intermediate metabolites, missing information on secondary metabolites, an overwhelming amount of kinetic information, much of which however is not useful for system identification, and multiple candidates for functional forms (other than power-laws).

All reported interactions of the primary metabolic pathway were considered and kinetic models were used to supplement the DFE. Dynamic flux profiles were derived for multiple wild-type experimental data sets with different initial glucose concentrations: 20mM, 40mM, and 80mM. When analyzed per DFE, these dynamic fluxes revealed unexpected and intriguing temporal patterns. To elucidate the mechanisms underlying these flux patterns a detailed kinetic model was fitted to one set of data (80mM glucose) by combining time-series data (of metabolites, cofactors and fluxes) with kinetic and regulatory information obtained from independent enzymatic studies. Subsequently, a qualitative functional analysis of the model was conducted to investigate the mechanisms that determined the peculiar trends observed in the DFE fluxes. In lieu of sensitivity analysis, the qualitative functional analysis provided insights into what controlling factors are likely to prevent faster glucose uptake into this pathway. The model was also tested

against another set of experimental data (40mM glucose) and the results of these analyses are presented here in this chapter.

Challenges in dynamic flux estimation

As is customary with DFE, I began with a detailed map for the glycolytic pathway in *L. lactis* (Figure 33) and with experimental data measured in the laboratory of Drs. Helena Santos and Ana Rute Neves with methods of *in vivo* nuclear magnetic resonance (NMR) under anaerobic conditions following a 20mM, 40 mM, or 80mM glucose bolus (Appendix B).

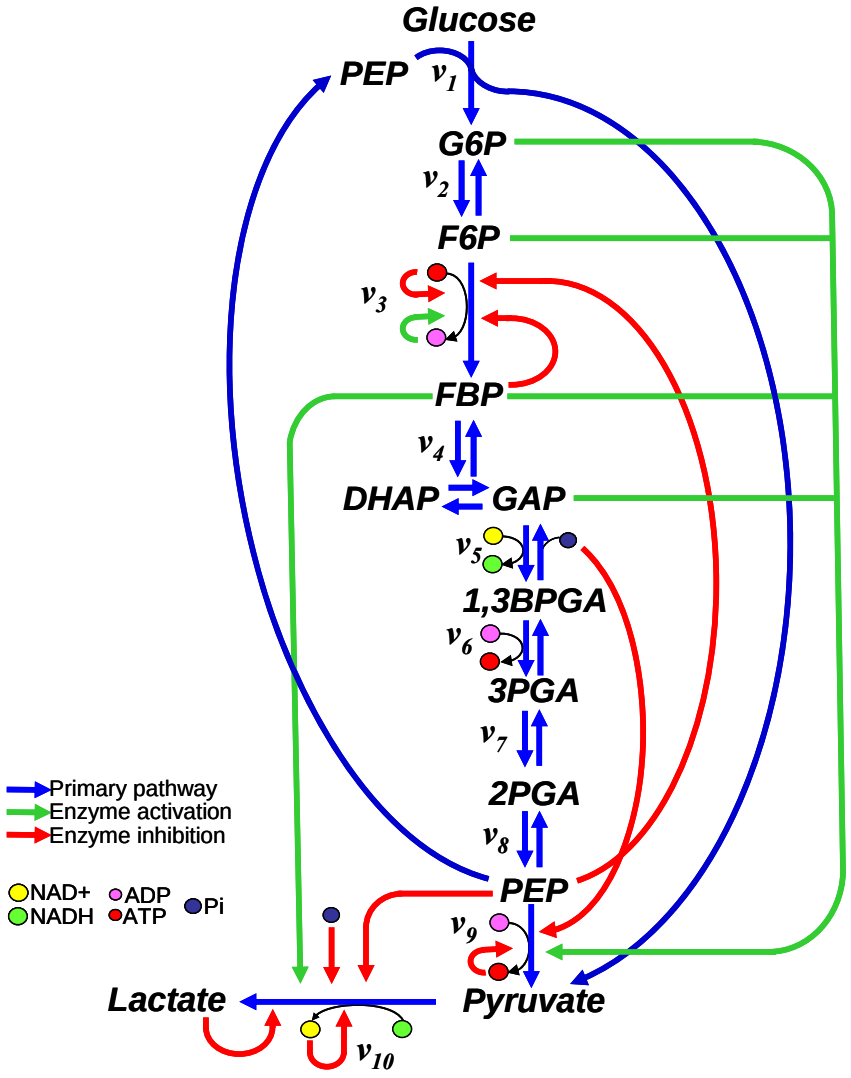


Figure 33: Schematic representation of glucose metabolism in *Lactococcus lactis*
Note in Figure 33 that the pathway does not include any secondary metabolites.

The NMR data provided measurements for only key metabolites in this pathway, which included FBP, 3PGA and PEP. When the mass balance was checked for each of the data sets, on average only 90% of carbon could be accounted for. This, however, is not attributable to measurement noise but rather is a limitation of the NMR technique, which detects metabolites only when they accumulate above a detection limit of about 2.5mM. Thus, the unaccounted mass is presumably distributed between the unobservable intermediate metabolites of the primary pathway (F6P, DHAP, GAP, 1,3BPGA, 2PGA and Pyruvate) as well as secondary metabolites. The secondary metabolites represent mass leaking out of glycolysis and entering various pathways (as discussed in Chapter 3; see Figure 22a), largely per catabolism of pyruvate, into mixed acids and other compounds such as aspartate, malate, succinate, acetoin and 2,3-butanediol. Moreover, 3PGA and PEP measurements are available only from the time after glucose is exhausted. There were, thus, three key challenges that prevented direct application of DFE to this pathway: (a) incomplete time-series data for primary metabolites (3PGA and PEP); (b) missing time-series data for intermediate metabolites (F6P, DHAP, GAP etc); and (c) missing information on secondary metabolites. To meet these challenges, I followed a phased approach to DFE whereby I successively addressed and resolved each of the three issues, beginning with the most simple pathway topology and later extending it to include aggregate leakage fluxes. The phases are discussed ahead.

Step 1: Addressing incomplete time-series data

The raw experimental data show (see Appendix B) that 3PGA and PEP measurements are available only for the time period after glucose is depleted. This is due to two reasons. PEP and 3PGA are not detected before addition of labeled glucose, because they are unlabeled, but after the glucose bolus and while glucose is present they are not detected because their concentration is below the detection limit. It is, however, safe to assume that the cell would have stored high levels of 3PGA and PEP during

starvation. Tandem experiments [34] have shown that 3PGA and PEP are in high concentrations before the addition of the second glucose bolus. These experiments provide guidance on what these time-series profile should look like while glucose is available. Consequently, artificial data points for 3PGA and PEP were introduced based on the tandem study. Likewise, since glucose 6-phosphate (G6P) was not measured for the specific bolus of 40mM and 80mM, it was adapted from the corresponding NMR experiment with 20 mM glucose bolus. Complete data on the key metabolites (G6P, FBP, 3PGA, PEP) were thus available, but data on other intermediate metabolites were not since they were practically below the detection limit of the specific NMR set-up (2.5mM).

Step 2: Selecting preliminary pathway topology

Before accounting for missing time-series data of intermediate metabolites, it was essential to make preliminary assumptions about the secondary metabolites or, in essence, about the pathway topology. Given the primary interest in understanding control of glycolysis, *i.e.*, the conversion of glucose into pyruvate, the secondary metabolites were ignored in the initial stages of flux computation. The pathway shown in Figure 33 is based on the assumption that there are no leakage fluxes and that all carbon mass is converted from glucose to lactate via this primary pathway. The lactate time-series data was hence adjusted to account for the remainder 10% mass, thus ensuring mass balance for the subsequent flux computations.

Step 3: Accounting for missing time-series data and estimating fluxes

Interestingly enough, each of the unobservable intermediate metabolites are products of a reversible step in the pathway and they exist in fast equilibrium with their respective precursor. Moreover, the kinetics of several of these enzymes has been characterized for *L. lactis* and the pertinent parameters are readily available [110]. These

include include phosphoglucose isomerase (pgi; flux $\boldsymbol{v_2}$); fructose 1,6-bisphosphate aldolase (fba; $\boldsymbol{v_4}$); phosphoglyceromutase (pgm; $\boldsymbol{v_7}$) and enolase (eno; $\boldsymbol{v_8}$). The flux through triosephosphate isomerase (tpi) is well known to be extremely fast and is considered to be always in equilibrium, implying that the rate of reaction is governed primarily by the equilibrium constant between GAP and DHAP.

By assuming a reversible Michaelis-Menten rate law for the net flux for each of the other enzymes, and using the parameter values obtained from literature [110, 111], the time-series data for all intermediate metabolites (F6P, DHAP, GAP and 2-PGA) were derived by solving the system sequentially rather than simultaneously [57]. Specifically, first the glucose influx into G6P pool (flux $\boldsymbol{v_1}$) was estimated using the slope estimates for the glucose time-series data. Subsequently, the efflux from G6P (flux $\boldsymbol{v_2}$) was estimated by solving the balance equation for rate of change of G6P by substituting time values of the slope of G6P time-series and flux $\boldsymbol{v_2}$. Next the time-series for F6P was derived by combining the kinetic model for PGI with the time series data on G6P and the flux estimates for v_2. Continuing in this fashion, the time-series data for fluxes and remainder of unobserved metabolites were computed sequentially.

The time-series data for 1,3BPGA however was artificially constructed because there are no known well-fitting kinetic rate-laws available for either glyceraldehyde 3-phosphate dehydrogenase (gapdh; $\boldsymbol{v_5}$) or phosphoglycerate kinase (pgk; $\boldsymbol{v_6}$). Thus, in all, 19 kinetic parameters were used from the literature for the 4 reversible Michaelis-Menten functions (for fluxes $\boldsymbol{v_2}$, $\boldsymbol{v_4}$, $\boldsymbol{v_7}$ and $\boldsymbol{v_8}$) (see model in Appendix C) and the time-series data were derived/constructed for 5 intermediate metabolites (F6P, DHAP, GAP, 1-3,BPGA, 2-PGA). (see Figure 34 for 80mM data set).

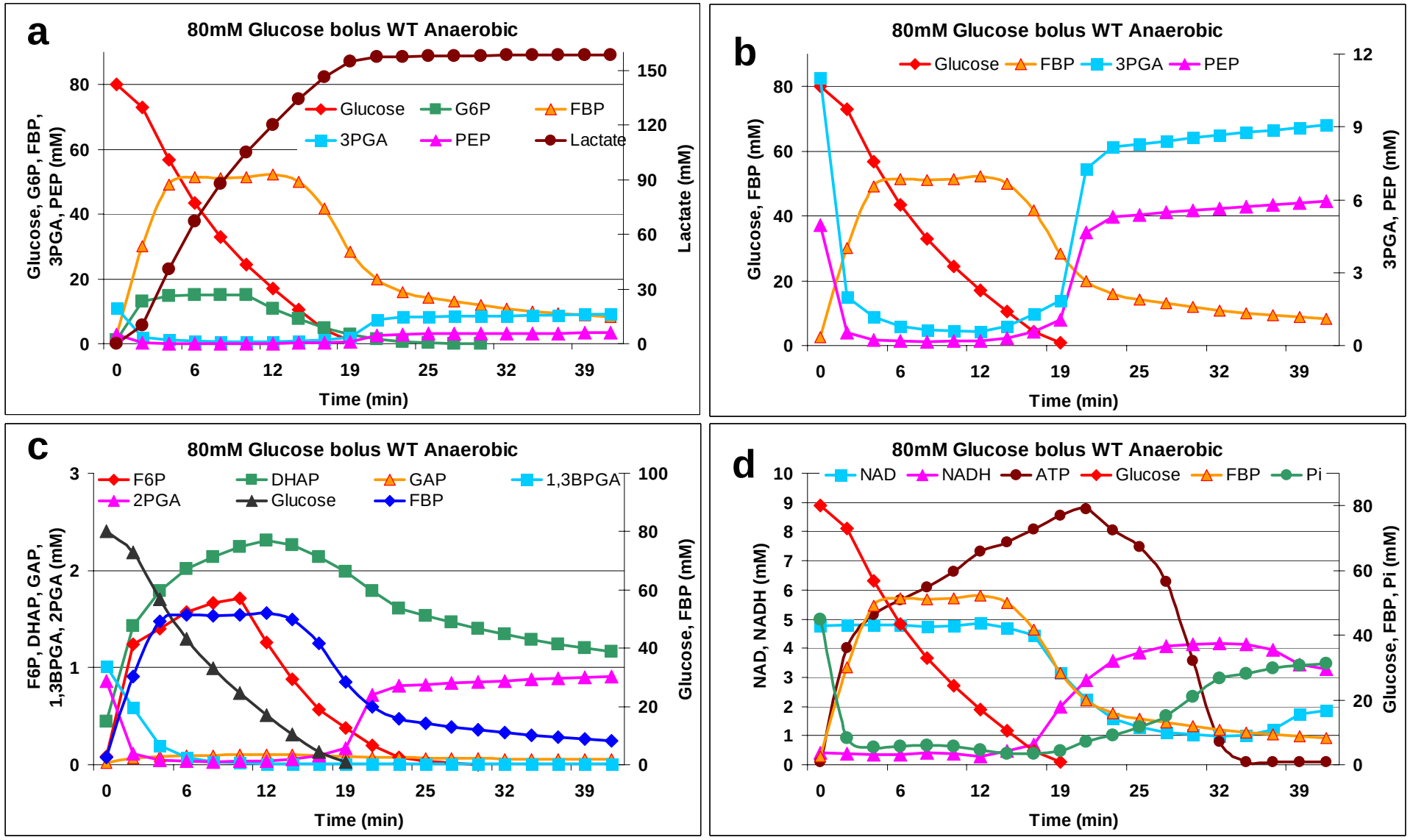


Figure 34: Smooth and derived time-series data based on glucose metabolism of *L.lactis* initiated with 80mM glucose bolus. (a), (b) Smoothed data for experimentally observable metabolites (c) Derived data for experimentally unobservable metabolites (d) Smoothed data for experimentally observable cofactors (NAD^+,NADH,ATP,Pi)

Step 4: Extending pathway topology using mass balance for co-factors

DFE offers the unique opportunity to estimate not only intracellular fluxes of the primary pathway but also fluxes of the coupled processes / enzymes, such as the ATPase and NADH-oxidase, which recycle the co-factors used in the primary pathway. Another advantage is that DFE facilitates the estimation of biochemical buffers, which are mostly a function of the experimental set-up rather than the biological system itself. For instance, the *in vivo* NMR data set, which quantifies both carbon metabolites and intracellular parameters (pH, ATP, Pi), had to be obtained from two distinct experiments where the cells were suspended in non-identical medium. The intracellular pools of intermediate metabolites were determined by ^{13}C-NMR using cells that were suspended in a 50mM potassium phosphate buffer (KPi; ph 6.5) whereas the intracellular parameters were monitored on-line by ^{31}P-NMR using cells suspended in a 50mM MES-NaOH buffer (pH 6.5). The Pi time-series data obtained using the second experiment obviously does not account for the phosphate rich medium used in the first experiment, which arguably served as a buffer to meet the demands of the metabolic pathway. The dynamic profile for this Pi buffer can be estimated by combining the time-series data from the two experiments with the DFE flux estimates.

Estimating the fluxes for coupled processes and/or buffers was a straightforward task within DFE. The ATPase flux was directly estimated by solving the following mass balance equation for ATP, using time-series data for the numerical DFE fluxes ($\boldsymbol{v_3}$, $\boldsymbol{v_6}$ and $\boldsymbol{v_9}$) and estimated slope of ATP time-series data.

$$\mathrm{ATPase} = -v3 + v6 + v9 - \dot{\mathrm{ATP}}$$

The estimated ATPase flux is shown in Figure 35a. Likewise, the Pi-buffer was estimated from the following balance equation for Pi (profile shown in Figure 35b).

$$\mathrm{Pi_{Buffer}} = \int \left(\dot{\mathrm{Pi}} + v5 - \mathrm{ATPase} \right) dt$$

The balance equation for NAD^+ was used to compute the aggregate leakage flux that oxidizes NADH to produce NAD^+. The mass accumulated from this aggregate flux is representative of all the secondary metabolites which are derived from pyruvate and produce NAD^+ along the pathways. These include pathways leading to metabolites such as alanine, aspartate, succinate and 2,3-butanediol. (In reality, aspartate and succinate are derived from oxaloacetate, which is derived from carboxylation of PEP and/or pyruvate, but the flux from PEP is not considered here).

$$Sec\_Metabolites_{NAD\_Regenerating} = \int \left(\dot{NAD} + v5 - LDH_{flux} \right) dt$$

There are, of course, other secondary metabolites derived from pyruvate as well, such as acetate, which do not involve NAD^+ regenerating pathways under anaerobic conditions and these metabolites are determined by establishing mass balance of carbon metabolites at the level of the entire pathway. Figure 35c shows the extensions to the primary pathway and Figure 35d shows accumulation of secondary metabolites derived from NAD^+ regenerating and NAD^+ independent fluxes.

Having extended the pathway to include aggregate leakage fluxes and coupled processes, the DFE flux profiles for all datasets (20mM, 40mM and 80mM) were analyzed for common trends and significant patterns. A detailed kinetic model was identified to fit the 80mM data and this model was analyzed qualitatively and tested against the 40mM dataset. These results are presented and discussed ahead.

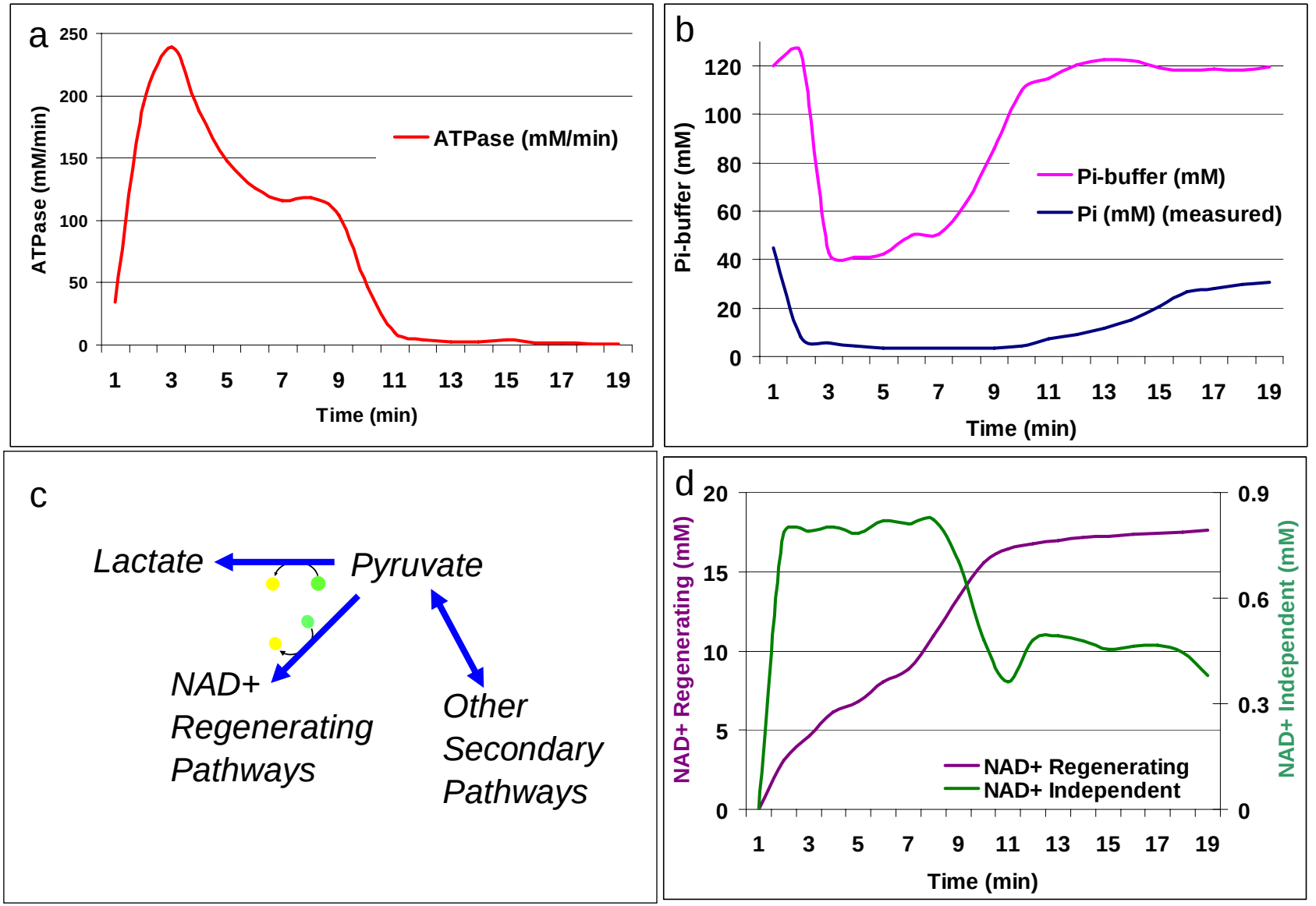


Figure 35: (a) Estimated ATPase profile **(b)** Estimated Pi-buffer and measured Pi

(c) Extensions to primary pathway **(d)** Estimated leakage mass, derived from pyruvate

DFE fluxes reveal unexpected temporal patterns

The flux profiles for all three datasets (20mM, 40mM, 80mM) are presented in Figure 36.When compared across the three data sets, these flux profiles exhibit similar trends and reveal three distinct phases underlying the observed wild-type dynamics:

A) ***Initial ramp-up and drive-down.*** Within the first 2 minutes, all fluxes in the system ramp up to a high value but before the end of the 3rd minute, all fluxes begin to slow down. This suggests global inhibition of all fluxes at/after 2-2.5 minutes including glucose transport. But this is counterintuitive: why would the bacterium, which has been starved of glucose, inhibit its glucose uptake when given a fresh bolus of sugar? Also note that all fluxes attain a higher value with lower glucose bolus than with higher glucose concentrations (max. flux value with 20mM glucose is 350mM/min whereas with 40mM and 80mM glucose it just above 250mM/min).

B) ***Quasi steady-state.*** This phase is predominantly visible in fluxes derived from 80mM data (Figure 36c). During the time-period of 3-10 minutes, all fluxes are found to be collectively decreasing and have the same value whereas all metabolites and cofactors in the system are found to be at approximately constant values during the same period (compare Figures 36c and 34). This is perplexing because the system has achieved steady state concentrations with decreasing flux values! This seems mathematically plausible but raises the questions of what regulatory mechanism enable this quasi-steady-state.

C) ***Reversal of flux ranks.*** At about 12 minutes (for 80mM data), the fluxes in the system depart from quasi steady state and there is a visible reversal in the rank of the fluxes compared to the ramp-up phase. This is true for 20mM and 40mM data as well; during the ramp-up phase, glucose transport is the fastest flux, but after the ramp-up and steady-state phase, the glucose transport

becomes the slowest flux. By the same token, the slowest flux from the first phase becomes the fastest flux in the third phase. What is confusing here is that the time point at which the glucose transport becomes the slowest does not correspond to the levels of glucose left for consumption. In 80mM it occurs at about the 12^{th} minute when glucose concentration is still more than 15mM!

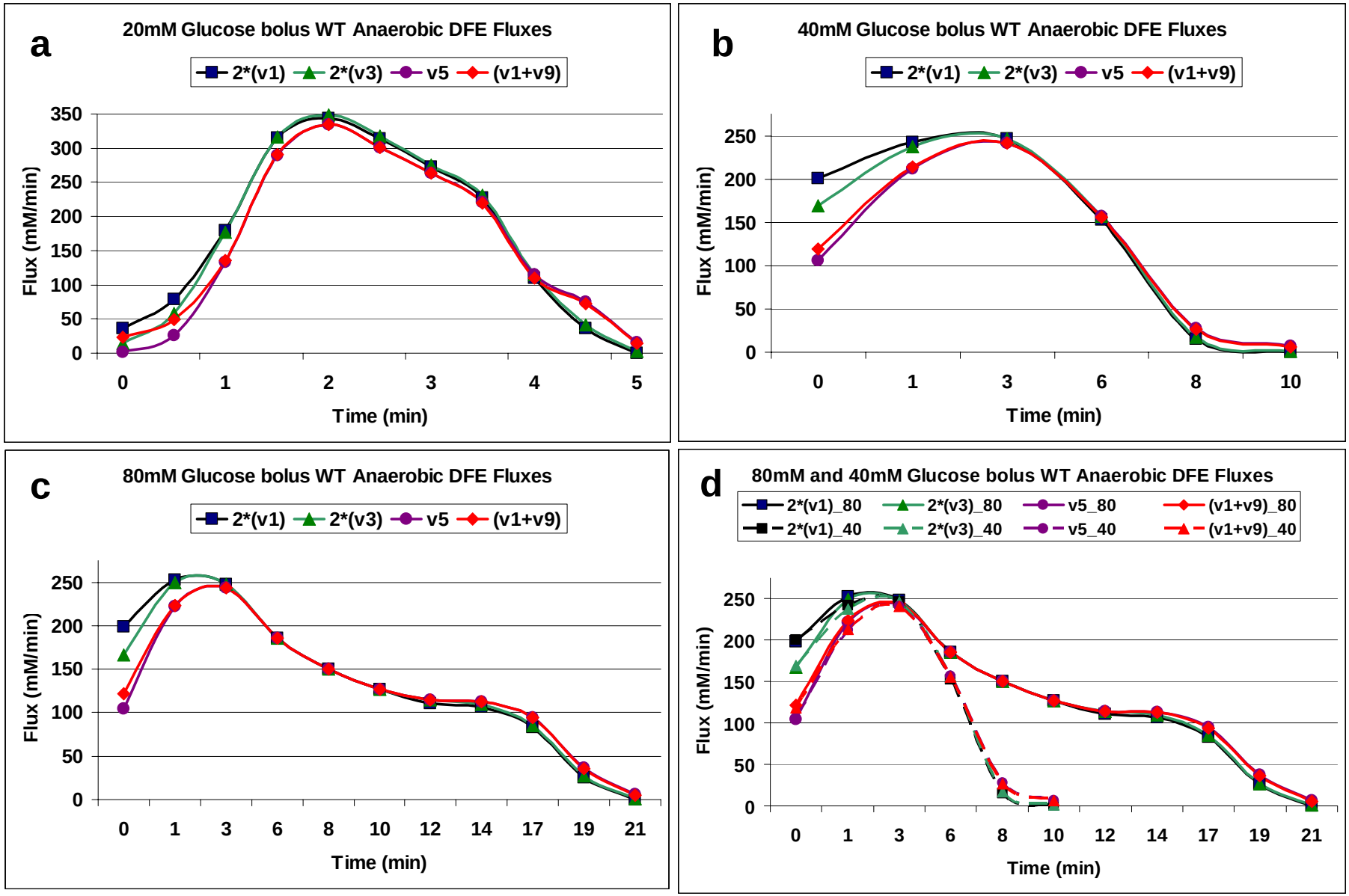


Figure 36: Dynamic flux profiles for (a) 20mM, (b) 40mM, and (c) 80mM wild-type data. Key representative fluxes are shown including glucose transport (v1), pfk (v3), gapdh (v5) and net efflux from PEP (v1+v9). Panel (d) compares fluxes derived from 40mM and 80mM glucose data sets.

Model Identification

The combined approach to estimating dynamic flux profiles resulted in not just the numerical flux estimates but also kinetic functions for 4 out of the total 10 fluxes that needed to be modeled in this system. The (symbolic) kinetic rate laws for the remainder of the fluxes were determined based on the regulatory information collected from the literature. The parameter values were optimized using the *lsqcurvefit* function in MATLAB where the objective function was to minimize the sum of least squared errors between the numerical fluxes, obtained from DFE, and values yielded by the kinetic function.

Glucose Transporter (flux $\boldsymbol{v_1}$)

Modeling glucose uptake presented a significant challenge because mechanisms of glucose transport regulation are not yet completely understood. As mentioned in the previous section, and shown in Figure 37, the glucose uptake was found to increase for the first 2-3 minutes and then continuously to decrease for the remainder of the time that glucose was available. Moreover, the glucose uptake flux was found to achieve faster speeds with lesser initial glucose. These observations led to several speculations and candidate models.

On the one hand, it was found that the observed slow-down of glucose transport could be modeled using an inhibition effect by a downstream intermediate metabolite such as G6P or FBP. It can be argued that the cells might employ such an inhibition mechanism to prevent high accumulation of phosphorylated metabolites which are toxic for the cells. However, there are no such reports of PTS activity inhibition by these metabolites.

On the other hand, it was found that inhibition by either of the substrates alone, glucose or PEP, could yield the same dynamic flux as obtained by DFE. But again, there

is no experimental evidence to confirm or disprove such inhibition. It stands to be argued that a third model incorporating inhibition by a fermentation product could very likely fit the data as well, but unless experimentally verified and proven, all these models will remain mere speculations.

These possible inhibition signals at the metabolic level are not the only factors to be taken into account. In addition, consequences of the experimental set-up should not be overlooked. For instance, the cell suspension is circulated through a 6-m-long loop connecting the bioreactor with the NMR tubes, where the actual measurements take place. Obviously, this transport causes a time delay. Nor must one ignore the fact that the glucose substrate is consumed not by a single cell but by a population of starving cells, which are likely to differ in membrane and transport properties that govern substrate uptake. It has been shown elsewhere [32] that if the uptake speed is more or less normally distributed among the cells, the resulting overall uptake characteristic is sigmoidal. Arguably the observed data describe glucose transport as a collective outcome of several of these processes.

Recently, Castro *et al*. [116] described all components involved in glucose transport in *L. lactis*. Their model included two PTS systems with distinct affinities for the two anomeric forms of glucose and a non-PTS permease. The model used here was directly adapted from that study:

$$V_1 = V_{max} \frac{\left(\frac{Glucose}{K_a}\right)^{n1}}{\left\{1+\left(\frac{Glucose}{K_a}\right)^{n1}\right\}} \frac{\left(\frac{PEP}{K_b}\right)^{n2}}{\left\{1+\left(\frac{PEP}{K_b}\right)^{n}\right\}} + V_{max} \frac{\left(\frac{Glucose}{K_a}\right)^{n1}}{\left\{1+\left(\frac{Glucose}{K_a}\right)^{n1}\right\}} \frac{\left(\frac{PEP}{K_b}\right)^{n2}}{\left\{1+\left(\frac{PEP}{K_b}\right)^{n}\right\}}$$

$$+ V_{max} \frac{\left(\frac{Glucose}{K_a}\right)^{n1}}{\left\{1+\left(\frac{Glucose}{K_a}\right)^{n1}\right\}} \frac{\left(\frac{ATP}{K_b}\right)^{n2}}{\left\{1+\left(\frac{ATP}{K_b}\right)^{n}\right\}}$$

Phoshpofructokinase (PFK) (flux v_3)

PFK catalyzes the transfer of a phosphoryl group from ATP to fructose-6-phosphate yielding fructose 1,6-bisphosphate:

$$\mathbf{F6P + ATP \rightarrow FBP + ADP + H^+}$$

The PFK reaction is essentially irreversible under cellular conditions, and it is the first "committed" step in the glycolytic pathway; G6P and F6P have other possible fates but FBP is targeted for glycolysis. In *Lactococcus*, PFK is reportedly activated by ADP and inhibited by ATP, PEP and FBP [52].

Inhibition of PKF by ATP is observed only when F6P is low (less than 0.6mM) or the ratio of ATP to F6P is greater than 1.Since in the given case study F6P was unobservable, it was assumed that F6P levels were very low in actuality and that ATP inhibition was thus present. Based on this information, the following rate law was formulated:

$$V3 = V_{max}\left(\frac{F6P}{K_{m1}\left(1+\frac{PEP}{K_i}\right)}\right)\left(\frac{ATP}{K_{m2}\left(1+\frac{ATP}{K_{iB}}\right)}\right)\frac{\left(\frac{F6P}{K_{m1}\left(1+\frac{PEP}{K_i}\right)}+\frac{FBP}{K_p}\right)^{n1-1}}{\left\{1+\left(\frac{F6P}{K_{m1}\left(1+\frac{PEP}{K_i}\right)}+\frac{FBP}{K_p}\right)^{n1}\right\}}\frac{\left(\frac{ATP}{K_{m2}\left(1+\frac{ATP}{K_{iB}}\right)}\right)^{n2-1}}{\left\{1+\left(\frac{ATP}{K_{m2}\left(1+\frac{ATP}{K_{iB}}\right)}\left(1+\frac{K_a}{ADP}\right)\right)^{n2}\right\}}$$

Glyceraldehyde 3-phosphate dehydrogenase (GAPDH) (flux v_5)

Oxidation of Ga3P to 1,3-bis-phosphoglycerate is catalyzed by Ga3PDH:

$$\mathbf{Ga3P + P_i + NAD^+ \leftrightarrow 1,3\text{-}PGA + NADH + H^+}$$

The reaction involves oxidation and phosphorylation of Ga3P by NAD^+ and Pi. The kinetics of this enzyme are reportedly very difficult to model not just for *L. lactis* but for yeast as well [112]. There is no clear consensus on the governing regulation which determines the wide range of activity exhibited by this enzyme. Different variants of Michaelis-Menten formalism have been proposed including reversible two-substrate two-product Michaelis-Menten functions with or without cooperativity, and some have included inhibition by adenine nucleotides. The model used for the present study was derived from the theoretical work of Hanekom [117]. The model is a specialized case derived from the general kinetic equations and is the following:

$$V5 = \left\{ V_f \left(\frac{GAP}{K_{s1}}\right)\left(\frac{Pi}{K_{s2}}\right)\left(\frac{NAD}{K_{s3}}\right) - V_r \left(\frac{BPGA}{K_{p1}}\right)\left(\frac{NADH}{K_{p2}}\right) \right\}$$

$$* \frac{\left\{ \left(\frac{GAP}{K_{s1}}\right)\left(\frac{Pi}{K_{s2}}\right) + \left(\frac{BPGA}{K_{p1}}\right) \right\}^{n1-1}}{\left\{ 1 + \left(\left(\frac{GAP}{K_{s1}}\right) + \left(\frac{BPGA}{K_{p1}}\right) \right)^{n1} + \left(\left(\frac{Pi}{K_{s2}}\right) + \left(\frac{BPGA}{K_{p1}}\right) \right)^{n1} + \left(\left(\frac{GAP}{K_{s1}}\right)\left(\frac{Pi}{K_{s2}}\right) + \left(\frac{BPGA}{K_{p1}}\right) \right)^{n1} - 2\left(\frac{Pi}{K_{s2}}\right)^{n1} \right\}}$$

$$* \frac{\left\{ \frac{NAD}{K_{s3}} + \frac{NADH}{K_{p2}} \right\}^{n2-1}}{\left\{ 1 + \left(\frac{NAD}{K_{s3}} + \frac{NADH}{K_{p2}} \right)^{n2} \right\}}$$

Phosphoglycerate kinase (PGK) (flux v_6)

PGK transfers the high energy phosphoryl group from the carboxyl group of 1,3-PGA to ADP, forming ATP and 3-PGA:

1,3-PGA + ADP ↔ 3-PGA + ATP

The kinetics for this enzyme was modeled based on the general derivation of a kinetic rate law for reversible two-substrate two-product Michaelis Menten with Hill effects [117].

$$V6 = \left\{ V_f \left(\frac{1,3BPG}{K_{s1}}\right).\left(\frac{ADP}{K_{s2}}\right) - V_r \left(\frac{3PGA}{K_{p1}}\right).\left(\frac{ATP}{K_{p2}}\right) \right\} \frac{\left\{\frac{1,3BPG}{K_{s1}} + \frac{3PGA}{K_{p1}}\right\}^{n1-1} \left\{\frac{ADP}{K_{s2}} + \frac{ATP}{K_{p2}}\right\}^{n2-1}}{\left\{1 + \left\{\frac{1,3BPG}{K_{s1}} + \frac{3PGA}{K_{p1}}\right\}^{n1}\right\}\left\{1 + \left\{\frac{ADP}{K_{s2}} + \frac{ATP}{K_{p2}}\right\}^{n2}\right\}}$$

Pyruvate Kinase (PK) (flux $\boldsymbol{v_9}$)

The last step in glycolysis is the transfer of the phosphoryl group from PEP to ADP, catalyzed by PK, which requires K^+ and either Mg^{2+} or Mn^{2+}:

PEP + ADP + H^+ → Pyruvate + ATP

In *Lactococcus*, pyruvate kinase is known to be inhibited by Pi and ATP while it is activated by FBP, G6P, DHAP, F6P, and GAP [52]. At low activator concentrations, the affinity of this enzyme for both PEP and ADP is reported to decrease (in the presence of saturated FBP concentration). Furthermore, Pi is reported to decrease the affinity for PEP and increase the FBP concentration required for half maximal velocity. Based on this regulatory information, the following model was constructed:

$$V9 = \frac{V_{max}(PEP)^{n1}}{\left\{\left(PEP\left(1+\frac{Pi}{K_i}\right)\right)^{n1}+\left(K_{s1}\left(1+\frac{K_{a1}}{FBP}\right)\left(1+\frac{K_{a2}}{DHAP}\right)\left(1+\frac{K_{a3}}{GAP}\right)\right)^{n1}\right\}}\frac{(ADP)^{n2}}{\left\{(ADP)^{n2}+\left(K_{s2}\left(1+\frac{K_{a4}}{G6P}\right)\left(1+\frac{K_{a5}}{F6P}\right)\left(1+\frac{ATP}{K_{p2}}\right)\right)^{n2}\right\}}$$

$$+\left\{\frac{V_{max}(PEP)^{n1}}{\left\{\left(PEP\left(1+\frac{Pi}{K_i}\right)\right)^{n1}+\left(K_{s1}\left(1+\frac{K_{a1}}{FBP}\right)\left(1+\frac{K_{a2}}{DHAP}\right)\left(1+\frac{K_{a3}}{GAP}\right)\right)^{n1}\right\}}\frac{(ADP)^{n2}}{\left\{(ADP)^{n2}+\left(K_{s2}\left(1+\frac{ATP}{K_{p2}}\right)\right)^{n2}\right\}}\right\}$$

The second term in the model for $\boldsymbol{v_9}$ above is a compensatory term that has a very low non-zero value even after G6P and F6P are depleted (unlike the first term which is reduced to zero). This term allows the system to channel mass from FBP into products downstream of pyruvate after glucose is depleted. In the absence of the second term, the $\boldsymbol{v_9}$ flux would be reduced to zero as soon as G6P or F6P are depleted which would lead to a very high accumulation of 3PGA and PEP in the system.

Lactate dehydrogenase (LDH) (flux v_{10})

For glycolysis to continue, NAD^+, which cells have in limited quantities, must be recycled after its reduction to NADH by Ga3PDH. In the absence of oxygen, NAD^+ is replenished by the reduction of pyruvate in an extension of the glycolytic pathway, either through homolactic or alcoholic fermentation. In the presence of oxygen, the there is an additional coupled process in which the reducing equivalents of NADH are oxidized by NADH oxidase (NOX). LDH specifically catalyzes the oxidation of NADH by pyruvate to yield NAD^+ and lactate:

Pyruvate + NADH ↔ Lactate + NAD^+

In *Lactococcus* LDH is activated by FBP but inhibited by high levels of intracellular PEP and Pi. Inhibition of LDH by Pi has been associated with an increase in the activation constant for FBP [52]. Based on this information, the flux through LDH was modeled as follows:

$$V13 = \left\{ V_f\left(\frac{Pyr}{K_{s1}}\right)\left(\frac{NAD}{K_{s2}}\right) - V_r\left(\frac{Lac}{K_{p1}}\right)\left(\frac{NAD}{K_{p2}}\right) \right\}$$

$$* \frac{\left(\frac{Pyr}{K_{s1}} + \frac{Lac}{K_{p1}}\right)^{n1-1}}{\left\{1 + \left(\frac{Pyr}{K_{s1}}\left(1 + \left(\frac{K_a}{FBP}\left(1 + \frac{Pi}{K_{i1}}\right)\right)^{n3}\right) + \frac{Lac}{K_{p1}}\left(1 + \frac{PEP}{K_{i1}}\right)\right)^{n1}\right\}} \frac{\left(\frac{NADH}{K_{s2}} + \frac{NAD}{K_{p2}}\right)^{n2-1}}{\left\{1 + \left(\frac{NADH}{K_{s2}} + \frac{NAD}{K_{p2}}\right)^{n2}\right\}}$$

Model Results

The parametric flux functions obtained from the model identification (Appendix C) fit the numerical fluxes computed using DFE (Figure 37). Thus, when integrated, this system of flux functions closely reproduces the observed experimental time courses of metabolites. The results for 80mM data set are shown in Figure 38.

The model fluxes demonstrate almost similar trends as those observed in DFE fluxes (Figure 39). There is the distinct initial ramp-up of fluxes, followed by a simultaneous steady decline of fluxes while glucose is available and the rank of the fluxes, from fastest to slowest, changes during the course of time (around the 12^{th} minute) even when there is abundant glucose available.

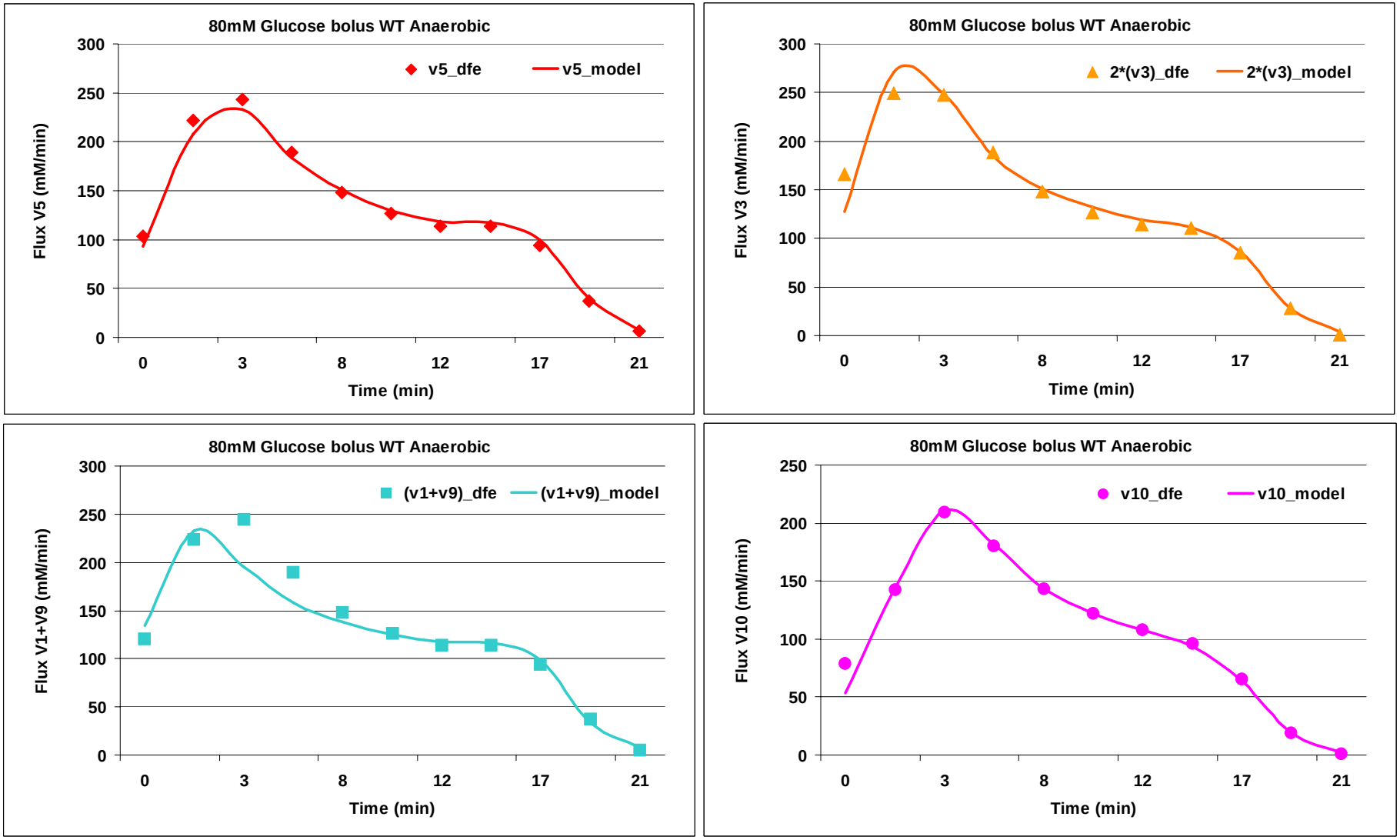


Figure 37: Model output (solid lines) contrasted with numerical fluxes computed using DFE for 80mM data set.

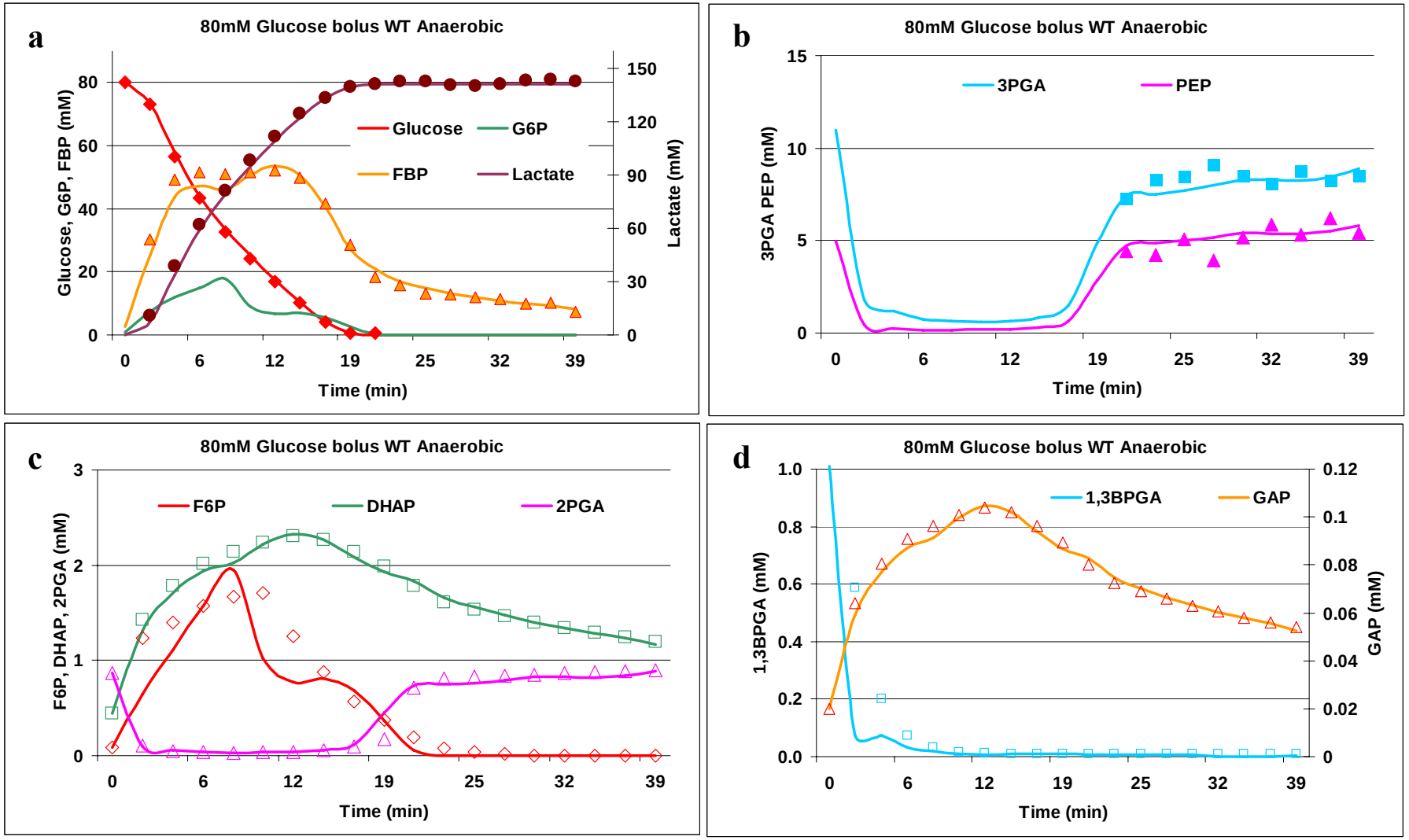


Figure 38: Model output (solid lines) contrasted with 80mM data set. Filled symbols in panels (a) and (b) represent experimental observations of metabolic data. Empty symbols in panels (c) and (d) represent artificial time-series data for experimentally unobservable intermediate metabolites

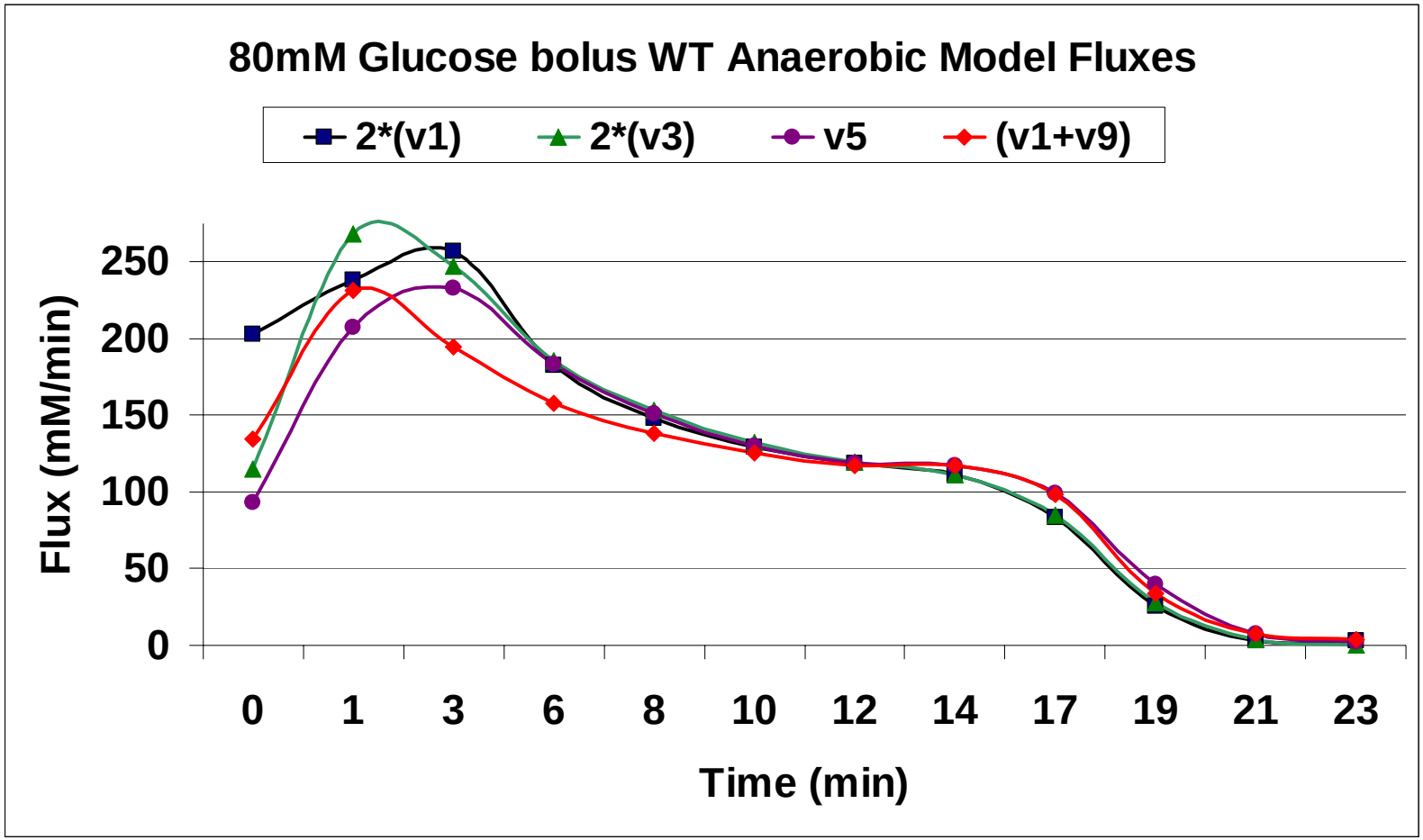


Figure 39: Intracellular fluxes derived from numerical integration of the detailed kinetic model for glycolysis in *L. lactis*. (see Appnedix C)

Model analysis: A case against sensitivity analysis

Having established a well-fitting model of glycolysis in the 80mM dataset, it is necessary to analyze whether any biologically meaningful information can be derived from this model. The first and foremost question typically asked is: which parameters have the greatest impact on biological outputs. The answer, of course, is obtained by conducting a sensitivity analysis where the parameters are ranked by a measure of the ratio of the fractional change induced in a biological variable of interest with respect to the fractional change induced in a parameter value. The general usefulness of this analytical tool is not in question but I do argue against its application to the kinetic model of glycolysis at hand for several reasons:

a. The system under investigation does not have a steady state while glucose is being consumed. The classical definition of sensitivity was developed for and applicable to systems in steady state only. Nonetheless, there are recent extensions and applications of sensitivity analysis to dynamical systems which include numerical computation of time-dependent sensitivities [118] [119], normalized sensitivities [120], and regional sensitivities [121]. But I cannot, as yet, use any of these numerical methods for the reason mentioned next.

b. The current kinetic model of glycolysis is still incomplete. Even though I have developed detailed kinetic models for each of the enzymes in the primary pathway, and have estimated coupled fluxes and hidden buffers in the system, I have not yet developed any models for the latter. As such, all co-factors (NAD^+, ATP, Pi) are still modeled using time-dependent uni-variate cubic splines. This makes it impossible to compute numerical sensitivities from this model because the co-factors will not change dynamically in response to a perturbation in any parameter. Alternatively, I could make the simple but naïve assumption that the co-factor profiles do not change significantly as

long as the parameters are perturbed within a few percent of their nominal value and thus compute sensitivities in that context. It can be argued that since these co-factors are consumed and produced in several processes other than the metabolic pathway, the organism would likely be robust to a 5% or 10% perturbation in parameters and hence the co-factor profiles do not change. But I have found, in the course of model identification for each of the enzymes, that they are very sensitive to minor fluctuations in co-factor profiles. Hence, I am not convinced that the exercise in sensitivity analysis on this incomplete model will be anything but futile.

c. Lastly, I do question the significance and/or the ability of sensitivity analysis to address the biological questions that interest my collaborators the most. The questions being: what controls glucose uptake? Can the organism/pathway be engineered to increase the speed of glucose consumption? If so, how? If not, why? Is the pathway already optimized to operate and consume glucose at the maximal speed possible?

In my view, the problem lies not in estimating dynamic sensitivities *per se* but in the belief that sensitivity analysis will successfully answer the question of what controls glucose uptake. In essence, the notion of sensitivity analysis mirrors and supports the practice of finding a "bottleneck" (in a sequential assembly of processes) with the hope that when the "bottleneck" is removed it will increase the flux through the entire system. And sensitivity analysis does just that: it measures fractional changes in a variable of interest in response to perturbation of a "single" parameter. I highly doubt that effecting such singular parametric changes experimentally will yield any significant improvements in glucose uptake. In fact, there is more than a decade long history of experimental research by several groups, including that of my collaborators, to alter the expression of each enzyme in the pathway,

one at a time, and none of these efforts have led to significant changes in glucose uptake by resting cells of *L.lactis*.

Moreover, if we were to experimentally perturb not one but the top two, three or five most sensitive parameters, there would be no guarantee that the individual parameter sensitivities would be preserved. It is likely that the parameter for which the system was observed to be sensitive, when perturbed alone, does not turn out to have the same affect on the system when two other parameters are changed simultaneously. There are two conventional ways to deal with this situation. One, we could compute and analyze numerical sensitivities for simultaneous changes induced in pairs or triplets or quadruplets of parameters. This, of course, leads to a combinatorial explosion of possibilities to analyze, considering that the current model has 84 parameters. Even if we start perturbing two parameters at a time and account for the fact that each parameter can be either increased or decreased from its estimated value, there are five different scenarios to be accounted for: values for both parameters are increased, decreased, or changed in opposite ways, or changed together randomly. This alone generates 5 x (83 x 84) x (2 x range of parameter values) possibilities to analyze. The total number of scenarios would thus be significantly higher when perturbing three or more parameters. The other-- and perhaps better-- alternative would be to set it up as an optimization problem: estimate minimum parametric variation that will maximize the speed of glucose uptake. This approach however is also challenged by the same perils as known in traditional parameter estimation and these have been researched and explored in depth by Torres and Voit [14].

What we thus need is a new approach to analyzing dynamical models, one that can effectively guide us to the solution of the biological questions of interest, in a more

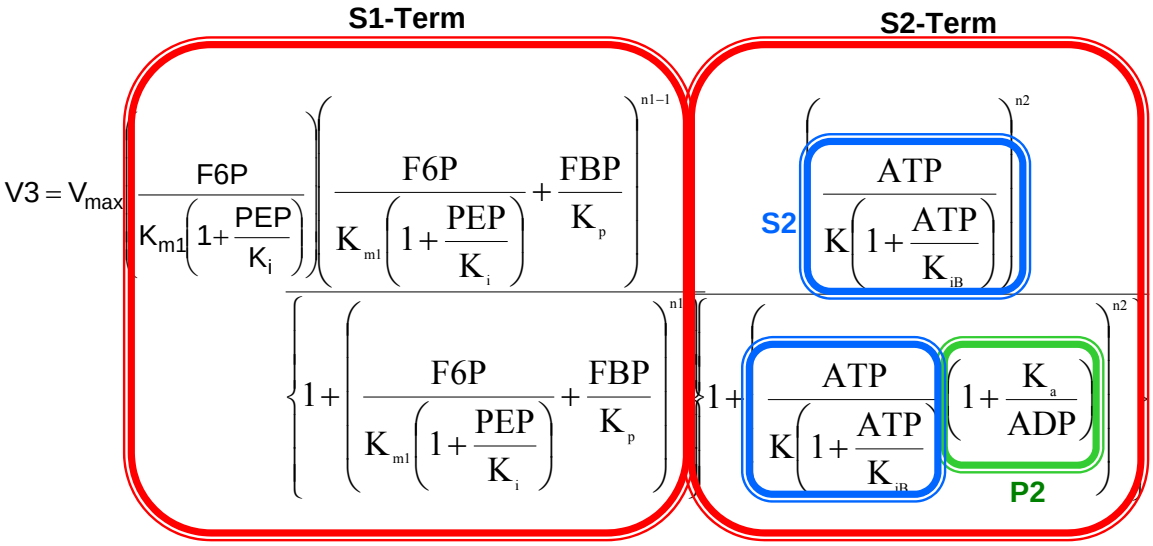
structured and less heuristic manner. A very simple and a novel approach, named "Qualitative Functional Analysis", is proposed here.

Qualitative Functional Analysis (QFA)

The motivation to devise this novel and simple analysis approach comes from an insistence to rephrase the very questions we asked when analyzing dynamical models. For the glycolytic model at hand, there remained two questions that demanded explanations: (a) what controls glycolysis; and (b) what are the mechanisms underlying the temporal patterns observed in DFE fluxes (Figure 36)?

By comparing the theoretical upper limit for each flux function (*i.e.*, its V_{max} value) with its maximum operating value (*i.e.*, the peak of the DFE flux profile), it was found that none of the flux functions were operating at their theoretical maximal limit. This led to the question: what prevents each flux in the pathway to achieve its theoretical maximum velocity? The hope was that by analyzing each flux separately, "Qualitative Functional Analysis" would answer this question, as discussed, in the following.

Consider the functional form for the phoshpofructokinase flux ($\boldsymbol{v_3}$) (shown below) which describes the kinetics for this enzyme as determined by two substrates (F6P and ATP), product inhibition (by FBP), competitive inhibition (by PEP), substrate self-inhibition (by ATP) and activation (by ADP). This function can be studied as a product of three terms: V_{max}, S1-term and S2-term. By substituting the time series data for each of the involved metabolites and co-factors in these terms, the lesser-valued function was identified to be S2-term (see Figure 40). This term was of interest because it suggested a clue to the question: what prevented flux $\boldsymbol{v_3}$ from operating at maximum velocity (V_{max})



Since the S2-term was identified to be the lesser valued term, it was further analyzed in terms of S2 and P2 (shown in blue and green boxes above). It turns out that the value of S2 was significantly high to the extent that the number 1 in the denominator could be easily ignored, and S2 thus be canceled between the numerator and the denominator, leaving only P2 in the denominator (see Figure 40). Thus, the observed values of the S2-term are primarily driven by the function P2.

In conclusion, QFA reveals that loss of activation of ADP results in very low values of P2, which make the S2-term the least valued term among V_{max}, the S1-term and the S2-term. Thus, loss of activation by ADP is most likely to prevent the flux $\boldsymbol{v_3}$ from operating at its theoretical maximal value.

Similar analyses were conducted for each step of the pathway, and the results of this analysis are shown in Appendix D.

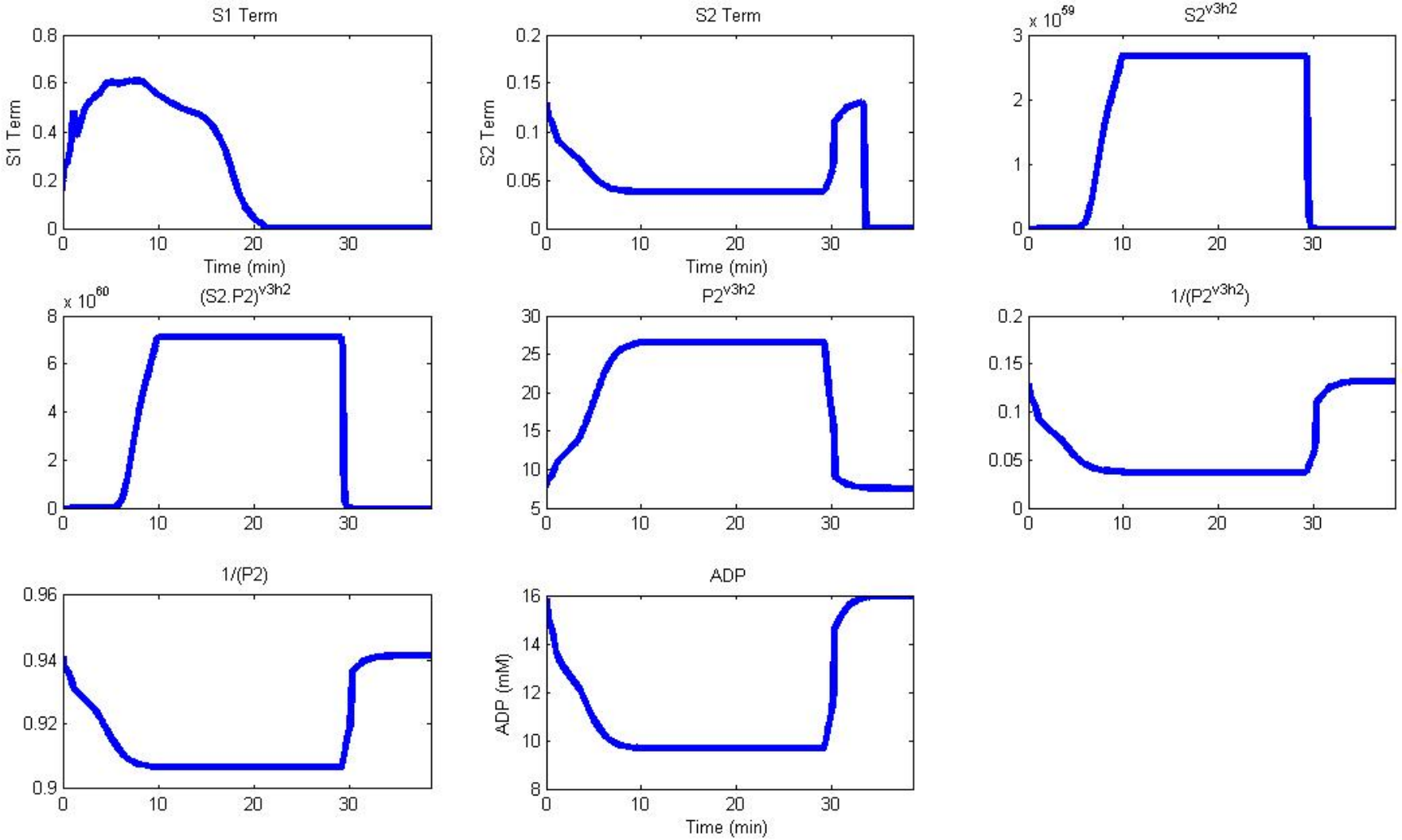


Figure 40: Qualitative Functional Analysis of flux v_3.

QFA: factors preventing faster glucose uptake

The following were observed after having analyzed each of the flux functions with QFA to investigate factors preventing faster fluxes at that specific step:

a. Loss of ADP activation prevents PFK ($\boldsymbol{v_3}$) from operating at a higher value
b. A higher NAD^+/NADH ratio would be required to drive the GAPDH ($\boldsymbol{v_5}$) flux higher
c. Reduction of ADP, which is a substrate for PGK ($\boldsymbol{v_6}$), keeps the flux at a lower level
d. Pi inhibition predominantly prevents PK ($\boldsymbol{v_9}$) flux from gaining higher speed
e. The glucose transport flux ($\boldsymbol{v_1}$) is primarily driven by both its substrates (Glucose and PEP) (S2-term) (see Figure D5 in Appendix D)

Model Validation

The norm in model development and analysis is to build a model using a training data set and then validate the model with a separate test data set that was not a part of the parameter optimization phase. In line with this practice, the kinetic model developed for wild type *L.lactis*, which reproduces the metabolic profile observed with 80mM glucose bolus under anaerobic conditions, was tested for 40mM glucose bolus under similar conditions. The model cannot be extrapolated to test for 20mM glucose because the necessary offline cofactor measurements are not available for that experiment. It was hoped that because the kinetic models were fitted to true intracellular fluxes, derived using DFE, that these functions would be easily and reliably extrapolated. Unfortunately, that is not the case.

Discussion

This case study showcases how DFE can be augmented with information from diverse sources to derive not just intracellular fluxes but also time-series data for unobserved processes that are coupled to observed processes. The model-free fluxes revealed unexpected patterns which raise interesting biological questions. When fitting functional forms for each of the fluxes, though DFE clearly avoided error compensation between fluxes, it seems that the kinetic models might have been "overfitted" for the intracellular fluxes of 80mM data set. It was attempted to re-optimize the kinetic models to fit to DFE fluxes for both 80mM and 40mM data set simultaneously but no good results were obtained at the time of writing this thesis. When trying to obtain a working model for 40mM data set alone, some of the kinetic parameter values obtained from the optimization were too high and unrealistic suggesting the need to alter the structure of the underlying function. This raises deeper questions about the assumption that a single model should be able to explain all datasets. Perhaps there are some additional factors, may be experimental or biological, which affect the regulation of the key enzymes that need to be accounted between these datasets. If not then the failure to extrapolate these functions suggests that the flux-substrate surface which involves complex regulation, such as what we have modeled here, cannot be reliably estimated from one data set alone. Future extensions of this work will have to consider fitting the model to several more replicates of data simultaneously to correctly approximate the flux surface in high dimensions.

Even though the current model could not extrapolate reliably, it could be used as a basis to gain insights into the underlying mechanisms of regulation. QFA revealed the factors that prevented each step from operating at its maximum theoretical velocity. From this local information about each step, it can be concluded that it should be possible to increase the glycolytic flux through the entire pathway by maintaining higher ratios of

ADP/ATP and NAD^+/NADH while counter-balancing the high inhibition of Pi with activation of high levels of FBP. This strategy might work provided the glycolytic flux is not further limited by the glucose uptake function itself which, as determined by QFA, seems to determined by both its substrates as well as non-metabolic, physiological factors. Even though this conclusion is based on the model derived from the 80mM dataset, which does not extrapolate well for other datasets, it will be interesting to see whether models derived for other datasets also lead to the same biological conclusions.

CONCLUDING REMARKS

In the course of this research, I believe we have significantly advanced our approach to system identification from time-series data. We gained deep insights into the issues beset with conventional approaches to parameter estimation. We isolated the true sources of error in these methods and proposed the novel approach of *Dynamic Flux Estimation (DFE)* which circumvents several of these issues. We have demonstrated the power of DFE both as a methodology and a framework which serves us well in a variety of non-ideal cases. But even before we arrive at a working kinetic model, which can be reliably extrapolated, DFE bears the unique ability to offer model-free insights into the underlying fluxes in the system. The fluxes derived with DFE, based either purely on measured time-series data or even when supplemented with a myriad of kinetic assumptions, hold within them complete information about local and global regulation of the system. As a computational technique, DFE adds immense power to the tools of experimental biology because it provides access into the unobservable state (enzyme activity) of the system from the measurable variables (metabolites) of the system. The mechanisms of regulation that we attempt to uncover from our study of metabolic pathways are in fact mechanisms acting at the level of enzymes. The metabolic system that we observe and quantify today are, in essence, recording the effects of a dynamical system of enzymes underplay. And DFE provides a reliable first means to translating the metabolic time-profiles into the underlying dynamic flux profiles. What remain to be developed are good approximations to model the system of fluxes. I firmly believe, with the growing number of time-series data that is becoming available in higher and higher quality, DFE is well positioned to lead the way in analyzing metabolic pathways.

APPENDIX A: PROOF-OF-CONCEPT MODEL

It is straightforward to construct different symbolic models for a metabolic system, such as the glycolytic pathway of interest here (Figure 22a, A1). In most formats, the underlying structure is given by the stoichiometry of the system. Thus, for each pool, a differential equation is set up that accounts for fluxes entering and leaving the pool. The equations for given case are shown in Figure A1.

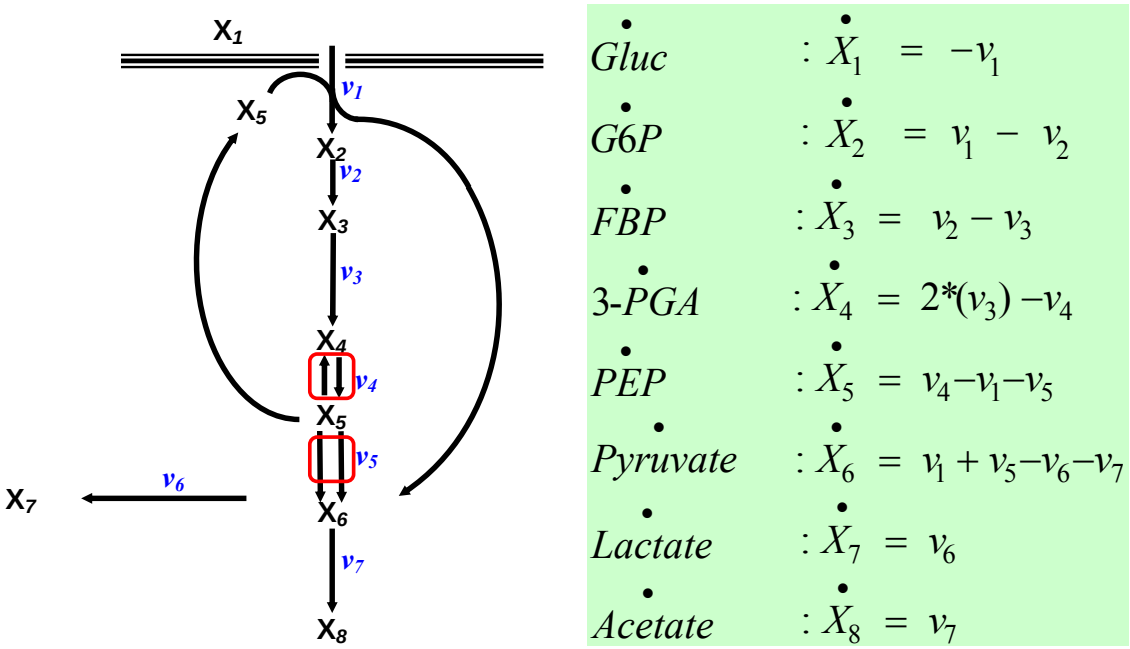


Figure A 1: Flux pattern in the *Lactococcus* pathway and corresponding, essentially assumption-free, mathematical representation of system of fluxes.

For many analytical purposes, it is convenient to reformulate this representation as a stoichiometric matrix equation of the following form

$$\begin{bmatrix} \dot{X}_1 \\ \dot{X}_2 \\ \dot{X}_3 \\ \dot{X}_4 \\ \dot{X}_5 \\ \dot{X}_6 \\ \dot{X}_7 \\ \dot{X}_8 \end{bmatrix} = \begin{bmatrix} -1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & -1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & -1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 2 & -1 & 0 & 0 & 0 \\ -1 & 0 & 0 & 1 & -1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 & -1 & -1 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \\ v_3 \\ v_4 \\ v_5 \\ v_6 \\ v_7 \end{bmatrix} \quad \text{.....(Eq. A1)}$$

At this point, mainstream models begin to differ. In standard stoichiometric and Flux Balance Analysis, it is assumed that the system is in steady state. Thus, the vector on the left-hand side is a vector of zeros. Furthermore, it is assumed that all fluxes are describable as constant flux rates, which is legitimate if the system is in a steady state. If all v_i are constant, Eq. A1 becomes a simple matrix equation that can be analyzed with methods of linear algebra.

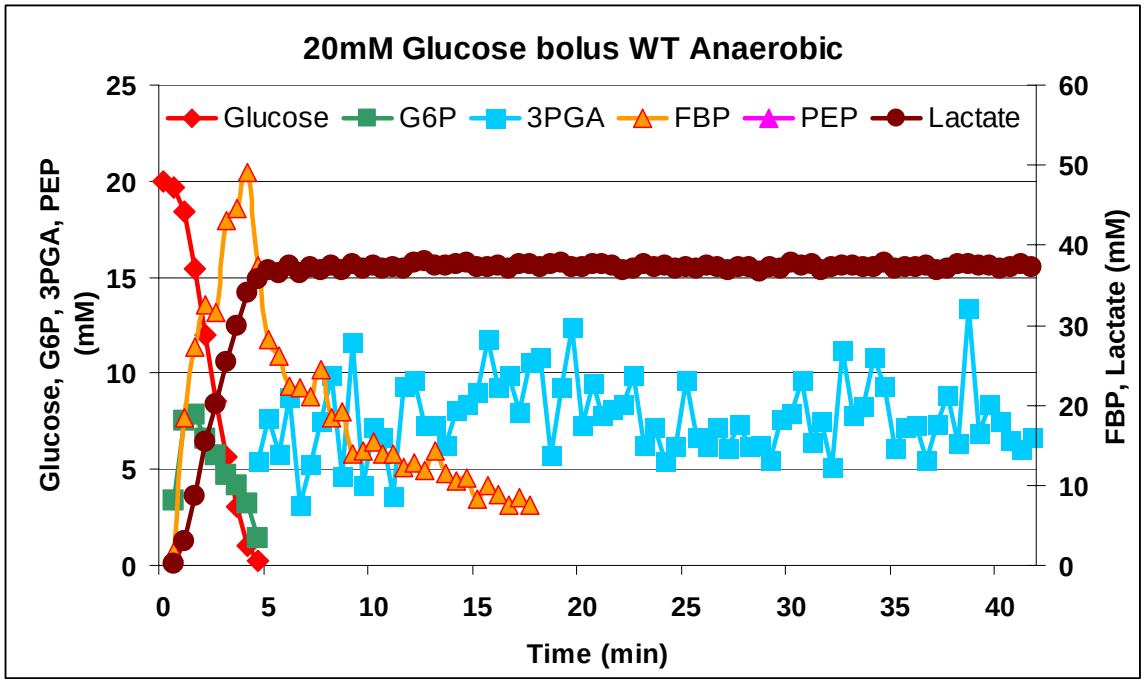
In the Generalized Mass Action format within Biochemical Systems Theory, the starting point is again the stoichiometric matrix equation. However, no assumption is made that the system is in a steady state. Furthermore, the fluxes are assumed to be functions of the system variables, and possibly other variables outside the system. As a consequence, the fluxes are functions of time-dependent variables and are therefore time dependent as well. Specifically, BST represents these fluxes as products of power-law functions that contain a rate constant and each contributing variable, raised to a real-valued kinetic order. In the given case (Chapter 3, Idealized Situation, Figure22), a numerical implementation of the glycolytic pathway in Lactococcus is given as shown in Figure. A2. The dynamic time courses of this particular model are shown in Figure 22b

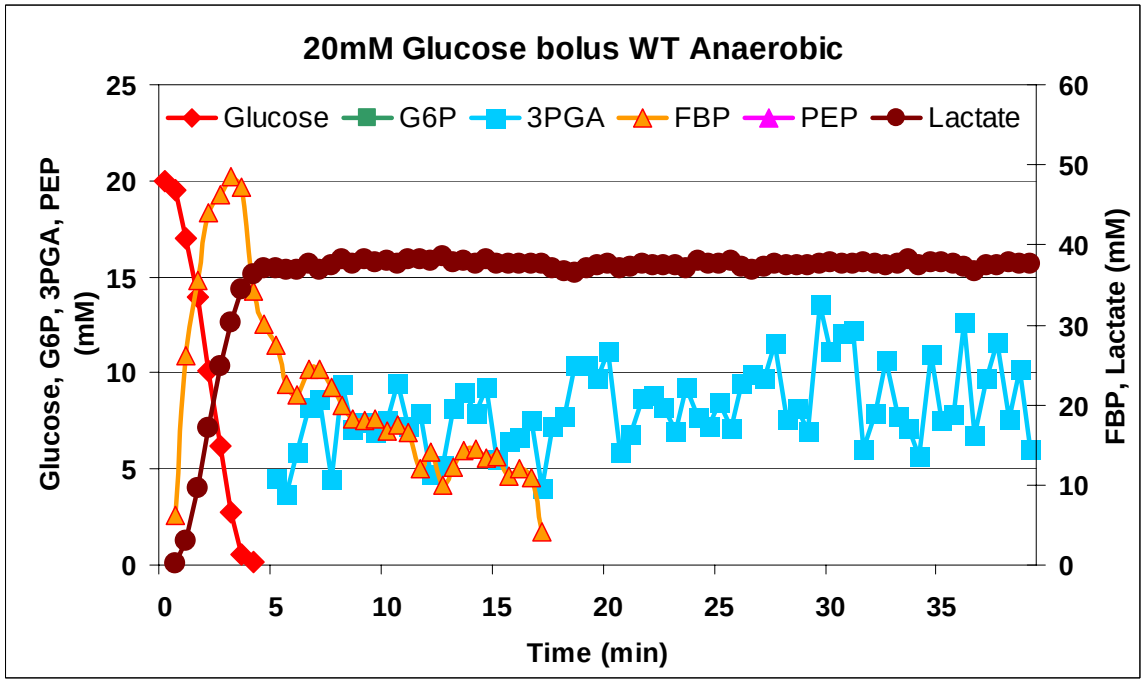
Flux	Power-law Flux Model
v_1	$9.47(\text{Glucose})^{0.4}(\text{PEP})^{0.81}$
v_2	$30.75(\text{G6P})^{0.74}(\text{ATP})^{0.4}$
v_3	$1.11(\text{FBP})^{0.88}(\text{Pi})^{0.01}$
v_4	$70.88(\text{PEP})^{0.43} - 31.97(\text{3PGA})^{0.32}$
v_5	$42.97(\text{PEP})^{0.53}(\text{FBP})^{1.33}(\text{Pi})^{-0.0001} + 8.08(\text{PEP})^{2.30}$
v_6	$100(\text{Pyruvate})^{0.46}(\text{FBP})^{1.04}$
v_7	$500(\text{Pyruvate})^{1.0}(\text{Pi})^{0.46}$

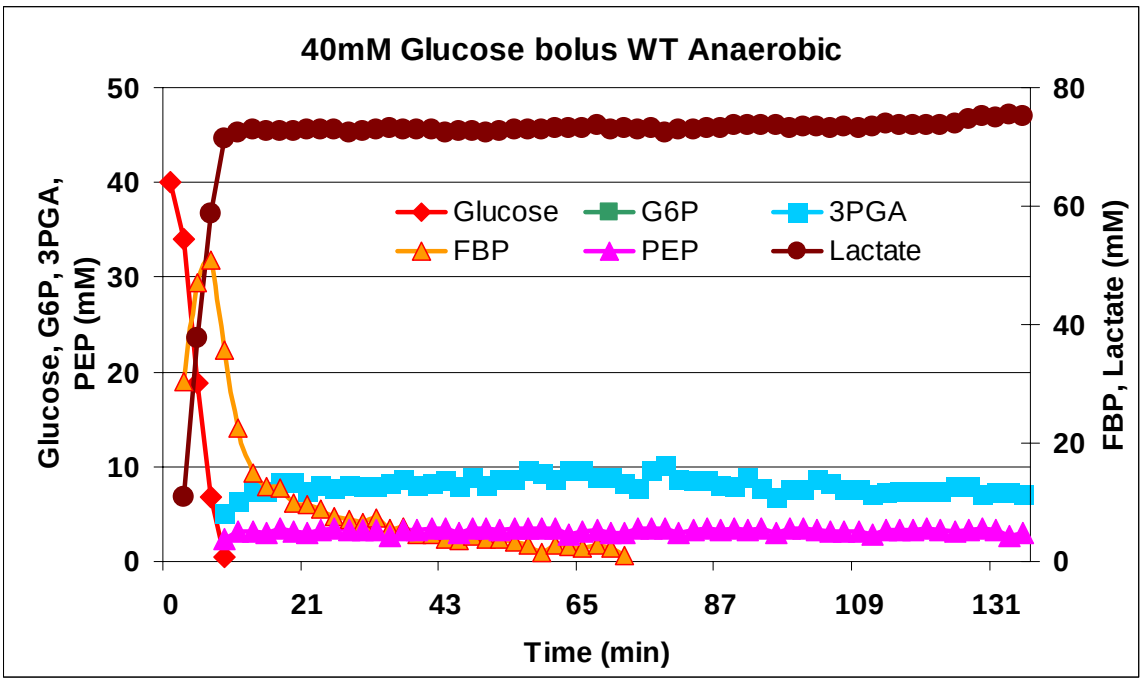
.

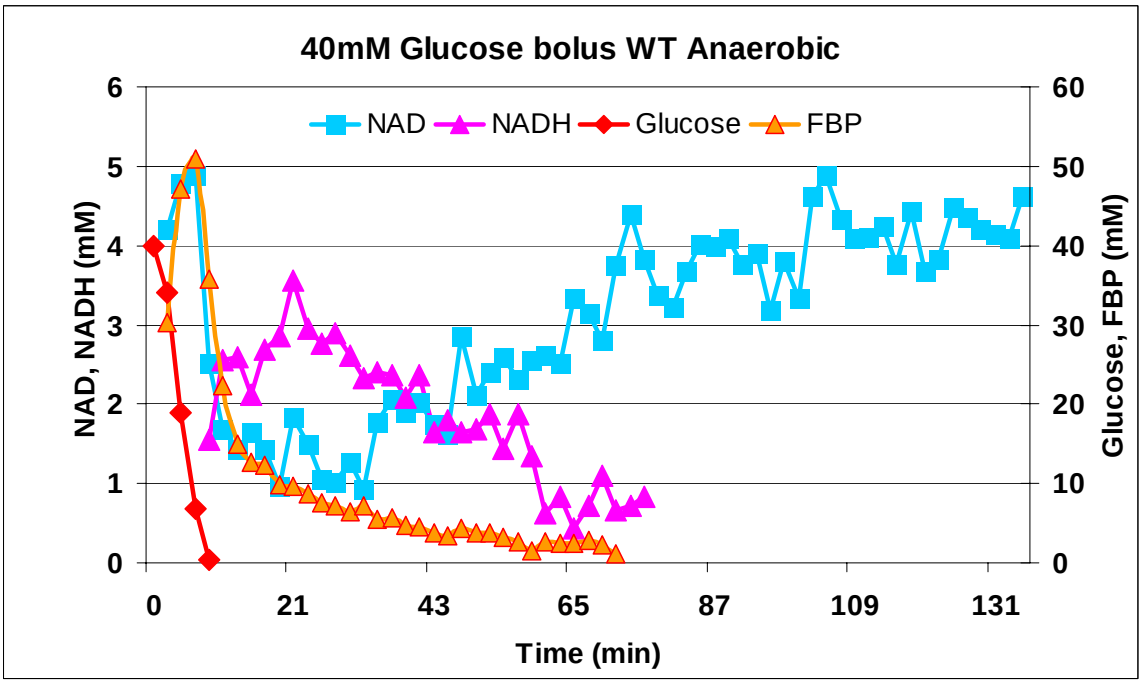
…….(Eq. A2)

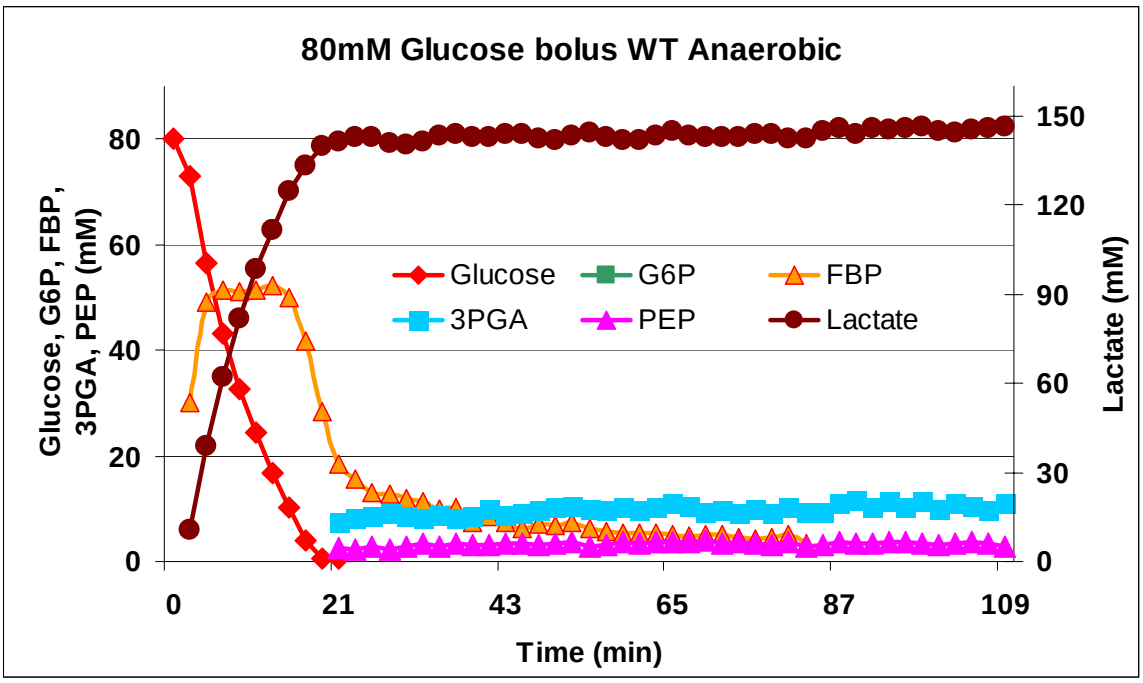
APPENDIX B: RAW EXPERIMENTAL DATA

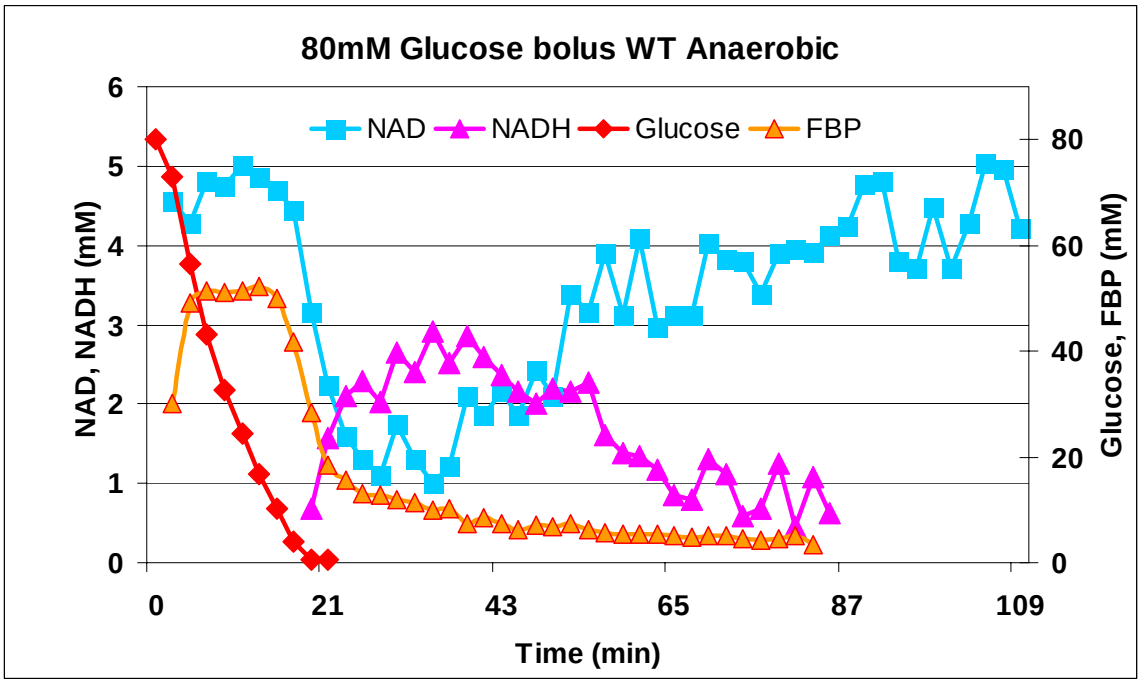












APPENDIX C: KINETIC MODEL FOR GLYCOLYSIS

Kinetic Functions

$$V_1 = V_{max} \frac{\left(\frac{Glucose}{K_a}\right)^{n1}}{\left\{1+\left(\frac{Glucose}{K_a}\right)^{n1}\right\}} \frac{\left(\frac{PEP}{K_b}\right)^{n2}}{\left\{1+\left(\frac{PEP}{K_b}\right)^{n}\right\}} + V_{max} \frac{\left(\frac{Glucose}{K_a}\right)^{n1}}{\left\{1+\left(\frac{Glucose}{K_a}\right)^{n1}\right\}} \frac{\left(\frac{PEP}{K_b}\right)^{n2}}{\left\{1+\left(\frac{PEP}{K_b}\right)^{n}\right\}}$$

$$+ V_{max} \frac{\left(\frac{Glucose}{K_a}\right)^{n1}}{\left\{1+\left(\frac{Glucose}{K_a}\right)^{n1}\right\}} \frac{\left(\frac{ATP}{K_b}\right)^{n2}}{\left\{1+\left(\frac{ATP}{K_b}\right)^{n}\right\}}$$

$$V2 = \frac{690.0585.\left(\frac{G6P}{1.5}\right) - 657.1986.\left(\frac{F6P}{0.2}\right)}{\left\{1+\frac{G6P}{1.5}+\frac{F6P}{0.2}\right\}}$$

$$V3 = V_{\max}\left(\frac{F6P}{K_{m1}\left(1+\frac{PEP}{K_i}\right)}\right)\left(\frac{ATP}{K_{m2}\left(1+\frac{ATP}{K_{iB}}\right)}\right)\frac{\left(\frac{F6P}{K_{m1}\left(1+\frac{PEP}{K_i}\right)}+\frac{FBP}{K_p}\right)^{n1-1}\left(\frac{ATP}{K\left(1+\frac{ATP}{K_{iB}}\right)}\right)^{n2-1}}{\left\{1+\left(\frac{F6P}{K_{m1}\left(1+\frac{PEP}{K_i}\right)}+\frac{FBP}{K_p}\right)^{n1}\right\}\left\{1+\left(\frac{ATP}{K\left(1+\frac{ATP}{K_{iB}}\right)}\left(1+\frac{K_a}{ADP}\right)\right)^{n2}\right\}}$$

$$V4 = \frac{935.6725.(\frac{FBP}{3}) - 162.9073.(\frac{DHAP}{0.13}).(\frac{GAP}{0.03})}{\left\{1+\frac{FBP}{3}+\frac{DHAP}{0.13}+\frac{GAP}{0.03}+(\frac{DHAP}{0.13}.\frac{GAP}{0.03})+(\frac{FBP}{3}.\frac{GAP}{0.23})\right\}}$$

$$V5 = \left\{ V_f \left(\frac{GAP}{K_{s1}}\right)\left(\frac{Pi}{K_{s2}}\right)\left(\frac{NAD}{K_{s3}}\right) - V_r \left(\frac{BPGA}{K_{p1}}\right)\left(\frac{NADH}{K_{p2}}\right) \right\}$$

$$* \frac{\left\{ \left(\frac{GAP}{K_{s1}}\right)\left(\frac{Pi}{K_{s2}}\right) + \left(\frac{BPGA}{K_{p1}}\right) \right\}^{n1-1}}{\left\{ 1 + \left(\left(\frac{GAP}{K_{s1}}\right) + \left(\frac{BPGA}{K_{p1}}\right) \right)^{n1} + \left(\left(\frac{Pi}{K_{s2}}\right) + \left(\frac{BPGA}{K_{p1}}\right) \right)^{n1} + \left(\left(\frac{GAP}{K_{s1}}\right)\left(\frac{Pi}{K_{s2}}\right) + \left(\frac{BPGA}{K_{p1}}\right) \right)^{n1} - 2\left(\frac{Pi}{K_{s2}}\right)^{n1} \right\}} \frac{\left\{ \frac{NAD}{K_{s3}} + \frac{NADH}{K_{p2}} \right\}^{n2-1}}{\left\{ 1 + \left(\frac{NAD}{K_{s3}} + \frac{NADH}{K_{p2}} \right)^{n2} \right\}}$$

$$V6 = \left\{ V_f \left(\frac{1,3BPG}{K_{s1}}\right).\left(\frac{ADP}{K_{s2}}\right) - V_r \left(\frac{3PGA}{K_{p1}}\right).\left(\frac{ATP}{K_{p2}}\right) \right\} \frac{\left\{ \frac{1,3BPG}{K_{s1}} + \frac{3PGA}{K_{p1}} \right\}^{n1-1}}{\left\{ 1 + \left\{ \frac{1,3BPG}{K_{s1}} + \frac{3PGA}{K_{p1}} \right\}^{n1} \right\}} \frac{\left\{ \frac{ADP}{K_{s2}} + \frac{ATP}{K_{p2}} \right\}^{n2-1}}{\left\{ 1 + \left\{ \frac{ADP}{K_{s2}} + \frac{ATP}{K_{p2}} \right\}^{n2} \right\}}$$

$$V7 = \frac{1111.1111.\left(\frac{3PGA}{1.2}\right) - 925.9259.\left(\frac{2PGA}{0.1}\right)}{\left\{ 1 + \frac{3PGA}{1.2} + \frac{2PGA}{0.1} \right\}};$$

$$V8 = \frac{1470.3501.\left(\frac{2PGA}{0.0372}\right) - 2722.8119.\left(\frac{PEP}{0.45}\right)}{\left\{ 1 + \frac{2PGA}{0.0372} + \frac{PEP}{0.45} \right\}}$$

$$V9 = \frac{V_{max}(PEP)^{n1}}{\left\{\left(PEP\left(1+\frac{Pi}{K_i}\right)\right)^{n1} + \left(K_{s1}\left(1+\frac{K_{a1}}{FBP}\right)\left(1+\frac{K_{a2}}{DHAP}\right)\left(1+\frac{K_{a3}}{GAP}\right)\right)^{n1}\right\}} \frac{(ADP)^{n2}}{\left\{(ADP)^{n2} + \left(K_{s2}\left(1+\frac{K_{a4}}{G6P}\right)\left(1+\frac{K_{a5}}{F6P}\right)\left(1+\frac{ATP}{K_{p2}}\right)\right)^{n2}\right\}}$$

$$+ \left\{ \frac{V_{max}(PEP)^{n1}}{\left\{\left(PEP\left(1+\frac{Pi}{K_i}\right)\right)^{n1} + \left(K_{s1}\left(1+\frac{K_{a1}}{FBP}\right)\left(1+\frac{K_{a2}}{DHAP}\right)\left(1+\frac{K_{a3}}{GAP}\right)\right)^{n1}\right\}} \frac{(ADP)^{n2}}{\left\{(ADP)^{n2} + \left(K_{s2}\left(1+\frac{ATP}{K_{p2}}\right)\right)^{n2}\right\}} \right\}$$

$$V13 = \left\{V_f\left(\frac{Pyr}{K_{s1}}\right)\left(\frac{NAD}{K_{s2}}\right) - V_r\left(\frac{LAC}{K_{p1}}\right)\left(\frac{NAD}{K_{p2}}\right)\right\} \frac{\left(\frac{Pyr}{K_{s1}} + \frac{LAC}{K_{p1}}\right)^{n1-1}}{\left\{1 + \left(\frac{Pyr}{K_{s1}}\left(1+\left(\frac{K_a}{FBP}\left(1+\frac{Pi}{K_{i1}}\right)\right)^{n3}\right) + \frac{LAC}{K_{p1}}\left(1+\frac{PEP}{K_{i1}}\right)^{n1}\right\}} \frac{\left(\frac{NADH}{K_{s2}} + \frac{NAD}{K_{p2}}\right)^{n2-1}}{\left\{1+\left(\frac{NADH}{K_{s2}} + \frac{NAD}{K_{p2}}\right)^{n2}\right\}}$$

Kinetic Parameters

v5KmGAP	0.0168
v5KmNAD	7.4247
v5KmPi	0.0000
v5KmBPGA	0.0000
v5KmNADH	0.1144
v5VmxFOR	599.9236
v5VmxREV	418.5004
v5h1	0.3148
v5h2	24.7483

v3KmF6P	0.0000
v3KmATP	0.1053
v3KiFBP	0.4628
v3KiPEP	0.0008
v3KiATP	647.1835
v3KaADP	0.9954
v3h1	0.0790
v3h2	33.4956
v3Vmx	3154.8279

v6KmBPGA	0.0000
v6Km3PGA	0.0001
v6KmADP	138.4104
v6KmATP	3.6761
v6VmxFOR	7796.1885
v6VmxREV	54.8435
v6h1	0.0729
v6h2	2.7232

v13KmPYR	0.1722
v13KmNADH	0.0002
v13KmLAC	9.6662
v13KmNAD	266.5506
v13KaFBP	0.0878
v13KiPi	0.0729
v13KiPEP	0.4077
v13VmxFOR	430.0131
v13VmxREV	0.1295
v13h1	1.1237
v13h2	48.4191
v13h3	0.4619

v9KmPEP	0.4049
v9KmADP	6.0154
v9KiPi	0.1274
v9KaFBP	3.8206
v9KaG6P	0.8146
v9KaF6P	0.0014
v9KaDHAP	0.0007
v9KaGAP	0.0013
v9KiATP	10.8326
v9h1	0.7359
v9h2	8.3415
v9Vmx	2623.3337

v9KmPEP	2.1092
v9KmADP	2.9372
v9KiPi	43.1158
v9KaFBP	0.0056
v9KaDHAP	0.926
v9KaGAP	0.0927
v9KiATP	0.0371
v9h1	15.6572
v9h2	0.4756
v9Vmx	10136.4683

V1Vmx	241.3592
v1KmGLU	280.0612
v1KmPEP	37.2885
v1h1	19.6019
v1h2	41.3087
a	2.2661
b	1.3015
c	0.4892

V1Vmx	247.1829
v1KmGLU	218.9962
v1KmPEP	0.3083
v1h1	1
v1h2	1
a	0.3149
b	0.0436
c	0.5042

V1Vmx	50
v1KmGLU	768.33
v1KmATP	2.9756
v1h1	4.2705
v1h2	19.1497
a	1.238
b	9.3437
c	1.8273

APPENDIX D: QUALITATIVE FUNCTIONAL ANALYSIS

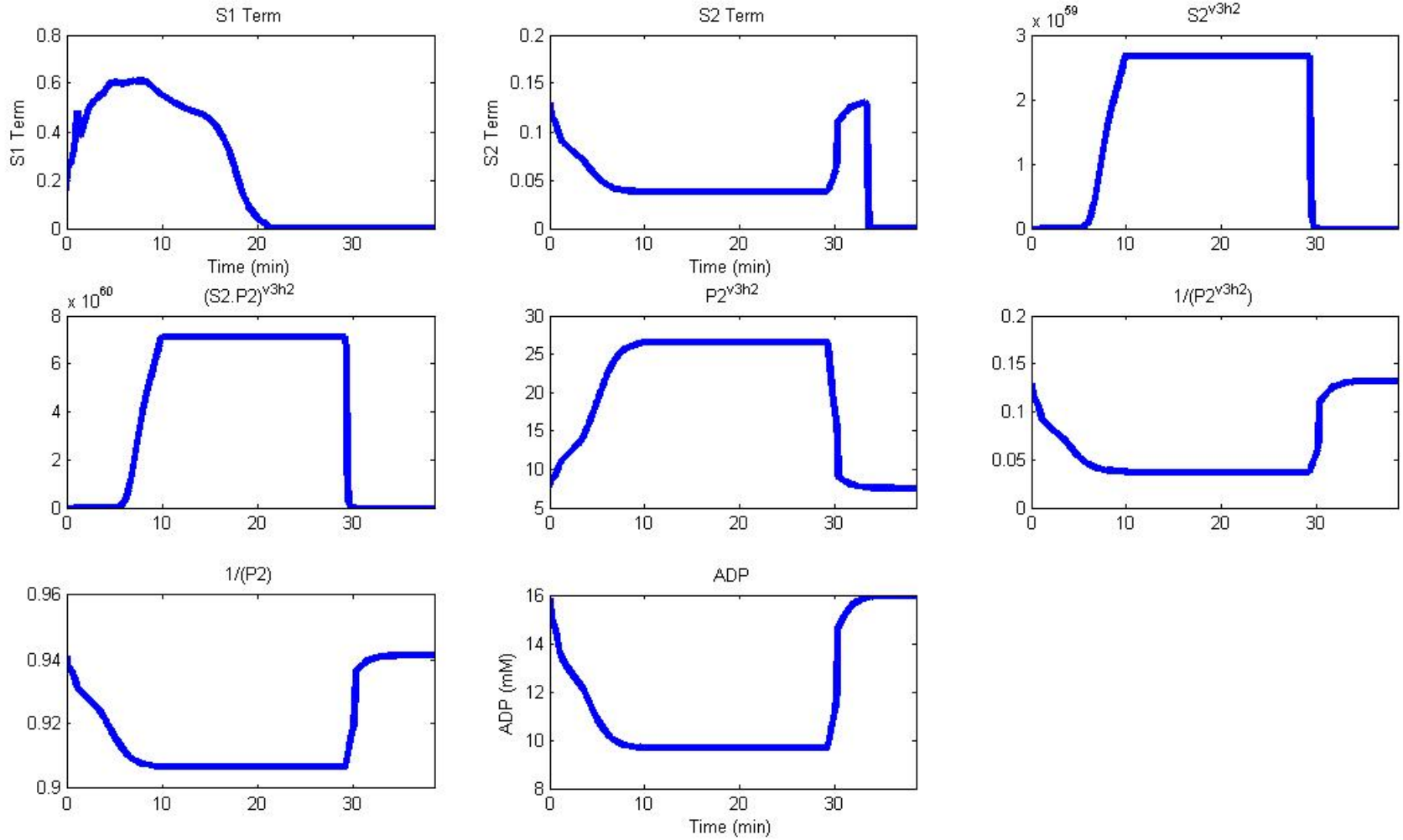


Figure D1: Qualitative Functional Analysis of flux v_3.

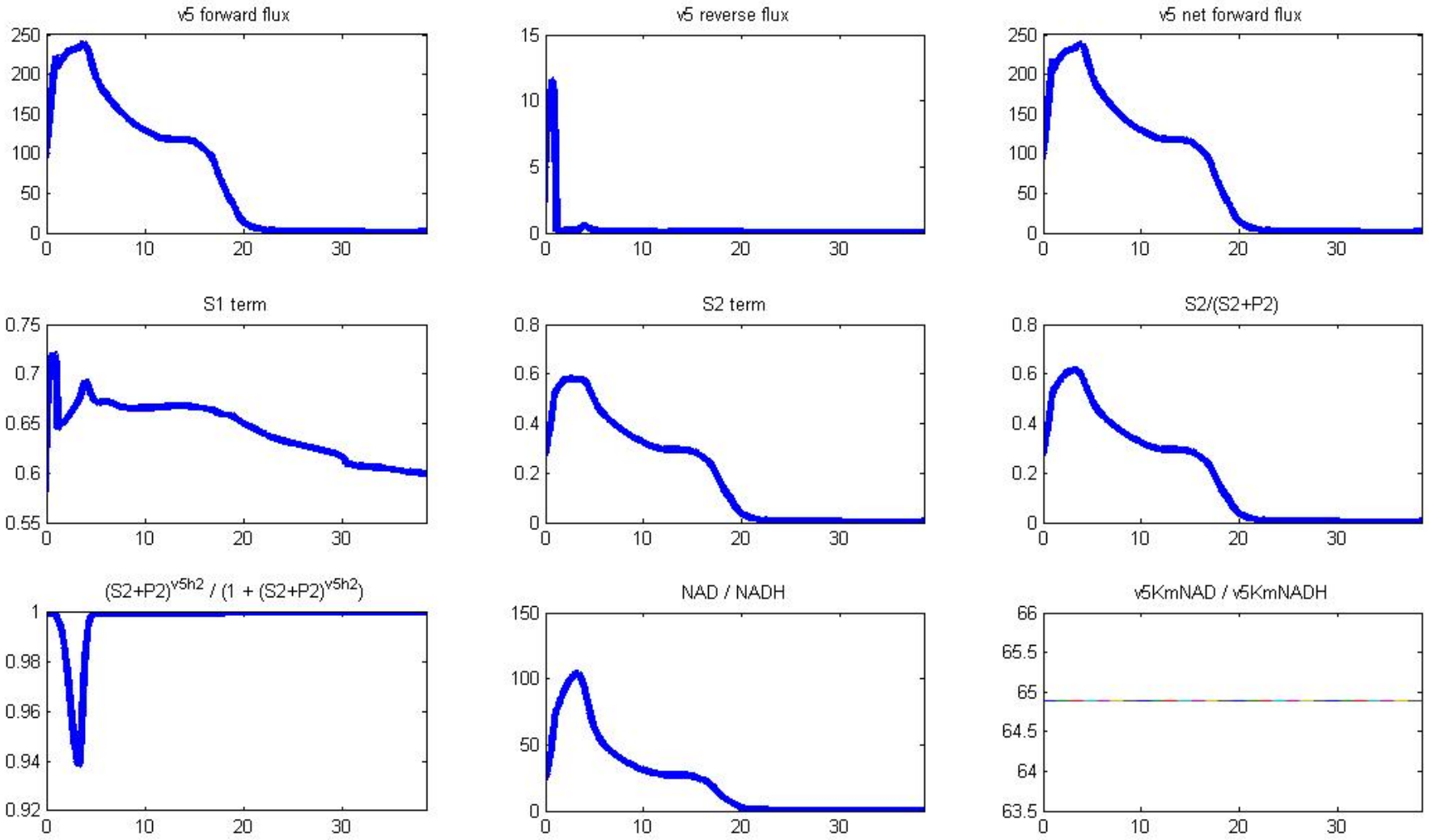


Figure D2: Qualitative Functional Analysis of flux v_5

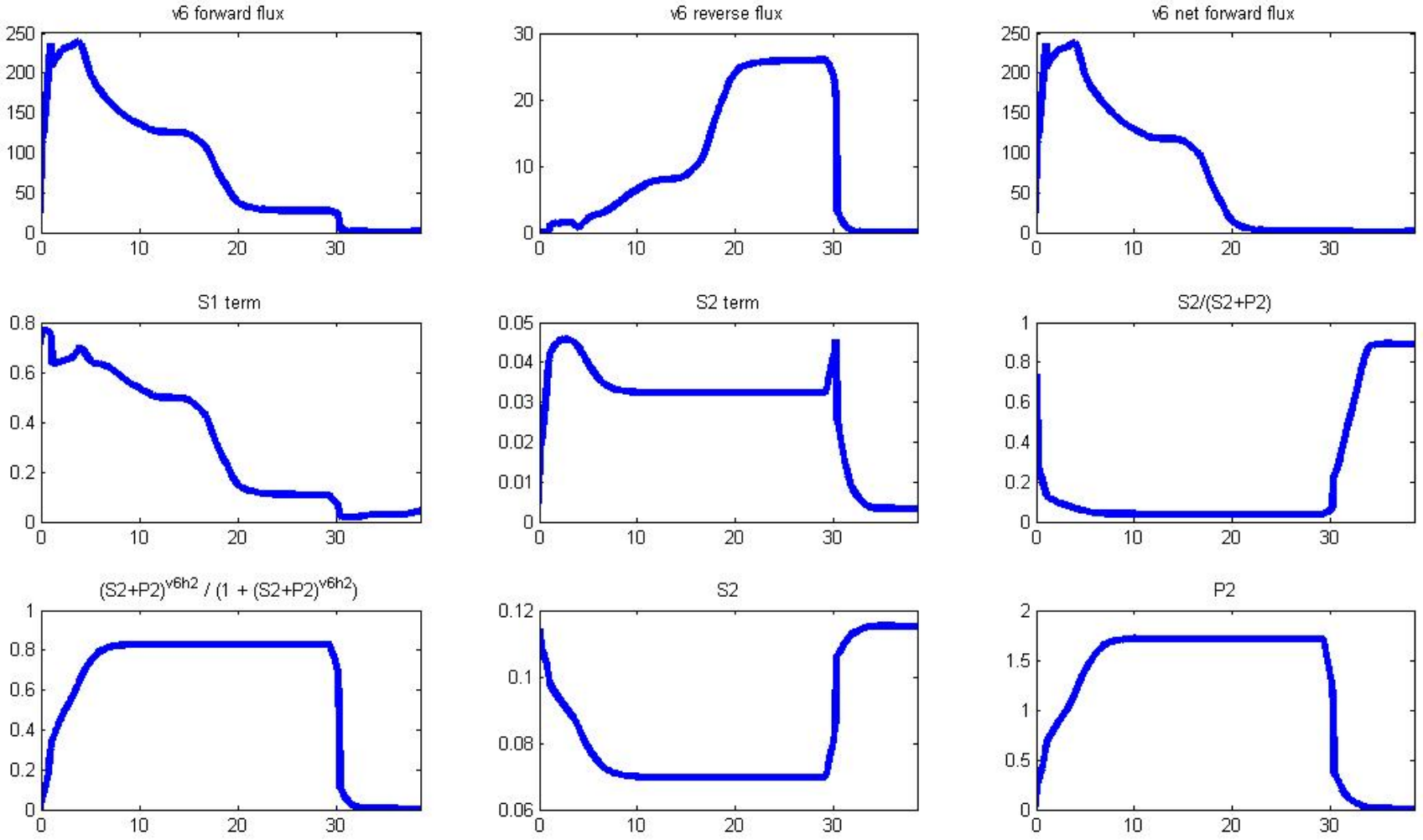


Figure D3: Qualitative Functional Analysis of flux v_6

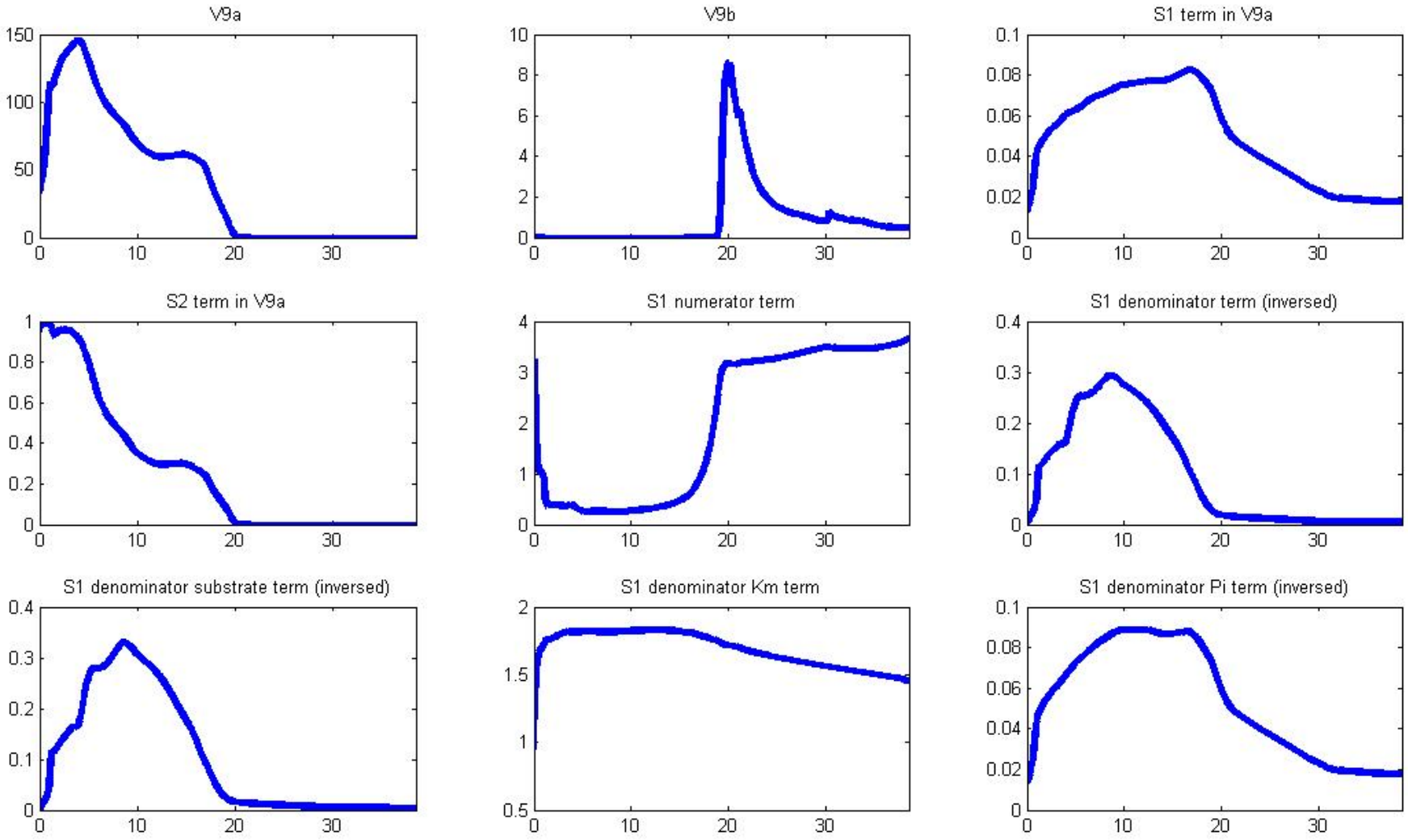


Figure D4: Qualitative Functional Analysis of flux v_9

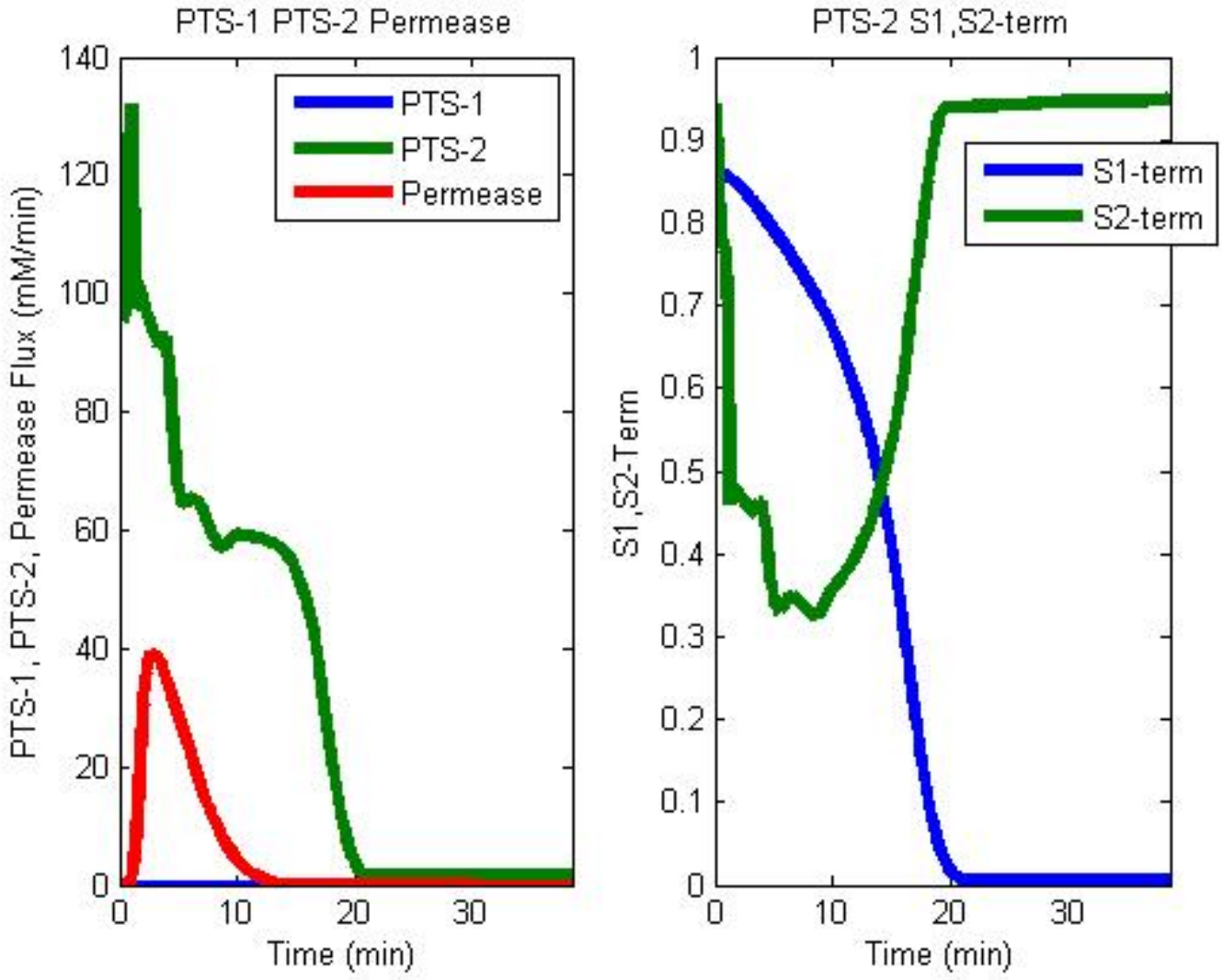


Figure D5: Qualitative Functional Analysis of flux v_1

REFERENCES

[1] M. A. SAVAGEAU, "The challenge of reconstruction". New Biol. **3**(2): p. 101-2, 1991

[2] G. GOEL, I.-C. CHOU and E. O. VOIT, "Biological systems modeling and analysis: A biomolecular technique of the twenty-first century". J. Biomol. Tech. **17**(4): p. 252-269, 2006

[3] Z. QI, G. W. MILLER and E. O. VOIT, "Computational systems analysis of dopamine metabolism". (submitted), 2008

[4] A. E. N. FERREIRA. "Power Law Analysis and Simulation Software". 2000; Available from: http://www.dqb.fc.ul.pt/docentes/aferreira/plas.html.

[5] P. MENDES. "Gepasi: Biochemical Simulation". 2004; Available from: http://www.gepasi.org/.

[6] J. SCHWACKE. "BSTLab: A Matlab Toolbox for Biochemical Systems Theory". Available from: http://bioinformatics.musc.edu/bstlab/.

[7] M. TOMITA. "BestKit". Available from: http://helios.brs.kyushu-u.ac.jp/~bestkit/.

[8] H. KURATA, K. MASAKI, Y. SUMIDA and R. IWASAKI, "CADLIVE dynamic simulator: direct link of biochemical networks to dynamic models". Genome Res. **15**(4): p. 590-600, 2005

[9] M. A. SAVAGEAU, "Biochemical systems analysis. I. Some mathematical properties of the rate law for the component enzymatic reactions". J Theor Biol. **25**(3): p. 365-9, 1969

[10] M. A. SAVAGEAU, "Biochemical systems analysis. II. The steady-state solutions for an n-pool system using a power-law approximation". J Theor Biol. **25**(3): p. 370-9, 1969

[11] M. A. SAVAGEAU, "Biochemical Systems Analysis: A Study of Function and Design in Molecular Biology. ". 1976: Advanced Book Program. Reading, MA: Addison-Wesley, .

[12] E. O. VOIT, "Canonical nonlinear modeling: S-system approach to understanding complexity". 1991: New York: Van Nostrand Reinhold.

[13] E. O. VOIT, "Computational Analysis of Biochemical Systems: A Practical Guide for Biochemists and Molecular Biologists. ". 2000: Cambridge University Press: New York.

[14] N. V. TORRES and E. O. VOIT, "Pathway Analysis and Optimization in Metabolic Engineering". 2002: New York: Cambridge University Press.

[15] R. CURTO, E. O. VOIT, A. SORRIBAS and M. CASCANTE, "Mathematical models of purine metabolism in man". Math Biosci. **151**(1): p. 1-49, 1998

[16] R. CURTO, E. O. VOIT, A. SORRIBAS and M. CASCANTE, "Validation and steady-state analysis of a power-law model of purine metabolism in man". Biochem J. **324 (Pt 3)**: p. 761-75, 1997

[17] R. CURTO, E. O. VOIT and M. CASCANTE, "Analysis of abnormalities in purine metabolism leading to gout and to neurological dysfunctions in man". Biochem J. **329 (Pt 3)**: p. 477-87, 1998

[18] M. A. SAVAGEAU and E. O. VOIT, "Recasting Nonlinear Differential-Equations As S-Systems - A Canonical Nonlinear Form". Mathematical Biosciences. **87**(1): p. 83-115, 1987

[19] F. SHIRAISHI and M. A. SAVAGEAU, "The tricarboxylic acid cycle in Dictyostelium discoideum. I. Formulation of alternative kinetic representations". J Biol Chem. **267**(32): p. 22912-8, 1992

[20] N. V. TORRES, "Modeling approach to control of carbohydrate-metabolism during citric-acid accumulation by aspergillus-niger .1. Model definition and stability of the steady-state". Biotechnology and Bioengineering. **44**(1): p. 104-111, 1994

[21] D. M. BATES and D. G. WATTS, "Nonlinear regression analysis and its applications.". Wiley series in probability and mathematical statistics. Applied probability and statistics. 1988, New York: Wiley.

[22] J. E. DENNIS, JR., D. M. GAY and R. E. WALSH, "An Adaptive Nonlinear Least-Squares Algorithm". ACM Trans. Math. Softw. **7**(3): p. 348-368, 1981

[23] M. MITCHELL, "An Introduction to Genetic Algorithms". Complex Adaptive Systems. 1996: MIT Press: Cambridge, MA.

[24] S. E. SELVAN, C. C. XAVIER, N. KARSSEMEIJER, J. SEQUEIRA, R. A. CHERIAN and B. Y. DHALA, "Parameter estimation in stochastic mammogram model by heuristic optimization techniques". IEEE Trans Inf Technol Biomed. **10**(4): p. 685-95, 2006

[25] R. H. J. M. OTTEN and L. P. P. GINNEKEN, "The Annealing Algorithm". 1989, Boston: Kluwer Academic Publishers.

[26] S. R. VEFLINGSTAD, J. ALMEIDA and E. O. VOIT, "Priming nonlinear searches for pathway identification". Theor Biol Med Model. **1**: p. 8, 2004

[27] G. GOEL, "Reconstructing Biochemical Systems: Systems Modeling and Analysis Tools for Decoding Biological Designs". 2008, Saarbrücken, Germany: VDM Verlag Dr. Müller.

[28] B. I. EPUREANU and H. S. GREENSIDE, "Fractal basins of attraction associated with a damped Newton's method". Siam Review. **40**(1): p. 102-109, 1998

[29] Z. KUTALIK, W. TUCKER and V. MOULTON, "S-system parameter estimation for noisy metabolic profiles using newton-flow analysis". IET Syst Biol. **1**(3): p. 174-80, 2007

[30] E. O. VOIT and J. ALMEIDA, "Decoupling dynamical systems for pathway identification from metabolic profiles". Bioinformatics. **20**(11): p. 1670-81, 2004

[31] E. O. VOIT and M. A. SAVAGEAU, "Power-law approach to modeling biological systems; III. Methods of analysis". J Ferment Technol. **60**(3): p. 223-241, 1982

[32] E. O. VOIT, S. MARINO and R. LALL, "Challenges for the identification of biological systems from in vivo time series data". In Silico Biol. **5**(2): p. 83-92, 2005

[33] I.-C. CHOU, H. MARTENS and E. O. VOIT, "Parameter estimation in biochemical systems models with alternating regression". Theor. Biol. Med. Model. **3**: p. 25, 2006

[34] A. R. NEVES, A. RAMOS, M. C. NUNES, M. KLEEREBEZEM, J. HUGENHOLTZ, W. M. DE VOS, J. ALMEIDA and H. SANTOS, "In vivo nuclear magnetic resonance studies of glycolytic kinetics in Lactococcus lactis". Biotechnol Bioeng. **64**(2): p. 200-12, 1999

[35] Y. SEKIYAMA and J. KIKUCHI, "Towards dynamic metabolic network measurements by multi-dimensional NMR-based fluxomics". Phytochemistry. **68**(16-18): p. 2320-9, 2007

[36] A. P. TEIXEIRA, S. S. SANTOS, N. CARINHAS, R. OLIVEIRA and P. M. ALVES, "Combining metabolic flux analysis tools and 13C NMR to estimate intracellular fluxes of cultured astrocytes". Neurochem Int. **52**(3): p. 478-86, 2008

[37] C. WITTMANN, "Fluxome analysis using GC-MS". Microb Cell Fact. **6**: p. 6, 2007

[38] C. YANG, Q. HUA and K. SHIMIZU, "Quantitative analysis of intracellular metabolic fluxes using GC-MS and two-dimensional NMR spectroscopy". J Biosci Bioeng. **93**(1): p. 78-87, 2002

[39] A. R. NEVES, W. A. POOL, J. KOK, O. P. KUIPERS and H. SANTOS, "Overview on sugar metabolism and its control in Lactococcus lactis - the input from in vivo NMR". FEMS Microbiol Rev. **29**(3): p. 531-54, 2005

[40] H. M. LEICESTER, "Development of Biochemical Concepts from Ancient to Modern Times". 1974, Cambridge, MA: Harvard University Press.

[41] E. J. CRAMPIN, M. HALSTEAD, P. HUNTER, P. NIELSEN, D. NOBLE, N. SMITH and M. TAWHAI, "Computational physiology and the Physiome Project". Exp Physiol. **89**(1): p. 1-26, 2004

[42] H. V. WESTERHOFF and B. O. PALSSON, "The evolution of molecular biology into systems biology". Nat Biotechnol. **22**(10): p. 1249-52, 2004

[43] L. VON BERTALALANFFY, "Der Organismus als physikalisches System betrachtet". Die Naturwissenschaften. **33**: p. 521-531, 1940

[44] N. V. TORRES, E. O. VOIT and C. H. ALCÓN, "Optimization of nonlinear biotechnological processes with linear programming. Application to citric acid production in Aspergillus niger". Biotechn Bioengin. **49**: p. 247-258, 1996

[45] A. E. N. FERREIRA, A. M. PONCES FREIRE and E. O. VOIT, "A quantitative model of the generation of N(epsilon)-(carboxymethyl)lysine in the Maillard reaction between collagen and glucose". Biochem J. **376**(Pt 1): p. 109-21, 2003

[46] E. O. VOIT, "Biochemical and genomic regulation of the trehalose cycle in yeast: review of observations and canonical model analysis". J Theor Biol. **223**(1): p. 55-78, 2003

[47] F. ALVAREZ-VASQUEZ, K. J. SIMS, L. A. COWART, Y. OKAMOTO, E. O. VOIT and Y. A. HANNUN, "Simulation and validation of modelled sphingolipid metabolism in Saccharomyces cerevisiae". Nature. **433**(7024): p. 425-30, 2005

[48] F. ALVAREZ-VASQUEZ, K. J. SIMS, Y. A. HANNUN and E. O. VOIT, "Integration of kinetic information on yeast sphingolipid metabolism in dynamical pathway models". J. Theor. Biol. **226**(3): p. 265-291, 2004

[49] R. ALVES, E. HERRERO and A. SORRIBAS, "Predictive reconstruction of the mitochondrial iron-sulfur cluster assembly metabolism: I. The role of the protein pair ferredoxin-ferredoxin reductase (Yah1-Arh1)". Proteins. **56**(2): p. 354-66, 2004

[50] M. VILELA, I.-C. CHOU, S. VINGA, A. T. VASCONCELOS, E. O. VOIT and J. S. ALMEIDA, "Parameter optimization in S-system models". BMC Syst Biol. **2**: p. 35, 2008

[51] A. RAMOS, A. NEVES, R. and H. SANTOS, "Metabolism of lactic acid bacteria studied by nuclear magnetic resonance". ANTONIE VAN LEEUWENHOEK INTERNATIONAL JOURNAL OF GENERAL AND MOLECULAR MICROBIOLOGY. **82**(1-4): p. 249-261, 2002

[52] A. R. NEVES, *Metabolic strategies to reroute carbon fluxes in Lactococcus lactis: kinetics of intracellular metabolite pools by in vivo Nuclear Magnetic Resonance.* Instituto de Tecnologia Química e Biológica. 2001, Portugal: Universidade Nova de Lisboa.

[53] E. O. VOIT, F. ALVAREZ-VASQUEZ and K. J. SIMS, "Analysis of dynamic labeling data". Math Biosci. **191**(1): p. 83-99, 2004

[54] R. CURTO, A. SORRIBAS and M. CASCANTE, "Comparative characterization of the fermentation pathway of Saccharomyces cerevisiae using biochemical systems theory and metabolic control analysis. Model definition and nomenclature.". Math Biosci. **130**: p. 25-50, 1995

[55] E. O. VOIT and A. E. N. FERREIRA, "Buffering in models of integrated biochemical systems". J Theor Biol. **191**: p. 429-438, 1998

[56] S. MARINO and E. O. VOIT, "An automated procedure for the extraction of metabolic network information from time series data". J. Bioinform. Comput. Biol. **4**(3): p. 665-691, 2006

[57] R. LALL and E. O. VOIT, "Parameter estimation in modulated, unbranched reaction chains within biochemical systems". Comput Biol Chem. **29**(5): p. 309-18, 2005

[58] V. L. CROW and G. G. PRITCHARD, "Purification and properties of pyruvate kinase from Streptococcus lactis". Biochim. Biophys. Acta,. **438**: p. 90-101, 1976

[59] W. MASON P, P. CARBONE D, A. CUSHMAN R and S. WAGGONER A, "The importance of inorganic phosphate in regulation of energy metabolism in Streptococcus lactis". Journal of Biological Chemistry. **256**(4): p. 1861-1866, 1981

[60] J. L. GALAZZO and J. E. BAILEY, "Fermentation pathway kinetics and metabolic flux control in suspended and immobilized *Saccharomyces cerevisiae*.". Enzyme Microb. Technol. **12**: p. 162-172, 1990

[61] E. O. VOIT, J. ALMEIDA, S. MARINO, R. LALL, G. GOEL, A. R. NEVES and H. SANTOS, "Regulation of glycolysis in Lactococcus lactis: an unfinished systems biological case study". Syst Biol (Stevenage). **153**(4): p. 286-98, 2006

[62] E. DENNIS J, M. GAY D and E. WELSCH R, "Adaptive Non-Linear Least Squares Algorithm". ACM Transactions on Mathematical Software. **7**(3): p. 348-368, 1981

[63] S. M. GOLDFELD, R. E. QUANT and H. F. TROTTER, "Maximization by quadratic hill-climbing". Econometrica. **34**(541-555), 1966

[64] Z. MICHAELEWICZ, "Genetic Algorithms + Data Structures = Evolution Programs". 1994, Berlin: Springer-Verlag.

[65] S. KIKUCHI, D. TOMINAGA, M. ARITA, K. TAKAHASHI and M. TOMITA, "Dynamic modeling of genetic networks using genetic algorithm and S-system". Bioinformatics. **19**(5): p. 643-650, 2003

[66] D. B. FOGEL, L. J. FOFEL and J. W. ATMAR. *Meta-evolutionary programming*. in *25th Asilomar Conference on Signals, Systems and Computers*: IEE Computer Society.

[67] S. KIMURA, K. IDE, A. KASHIHARA, M. KANO, M. HATAKEYAMA, R. MASUI, N. NAKAGAWA, S. YOKOYAMA, S. KURAMITSU and A. KONAGAYA, "Inference of S-system models of genetic networks using a cooperative coevolutionary algorithm". Bioinformatics. **21**(7): p. 1154-63, 2005

[68] G. GOEL, I. C. CHOU and E. O. VOIT, "Biological systems modeling and analysis: a biomolecular technique of the twenty-first century". J Biomol Tech. **17**(4): p. 252-69, 2006

[69] G. GOEL, *Biochemical Systems Toolbox*. Bioengineering. 2006, Atlanta: Georgia Institute of Technology. 136.

[70] D. Y. CHO, K. H. CHO and B. T. ZHANG, "Identification of biochemical networks by S-tree based genetic programming". Bioinformatics. **22**(13): p. 1631-40, 2006

[71] T. DAISUKE and P. HORTON, "Inference of scale-free networks from gene expression time series". J Bioinform Comput Biol. **4**(2): p. 503-14, 2006

[72] O. R. GONZALEZ, C. KUPER, K. JUNG, P. C. NAVAL, JR. and E. MENDOZA, "Parameter estimation using Simulated Annealing for S-system models of biochemical networks". Bioinformatics. **23**(4): p. 480-6, 2007

[73] K.-Y. KIM, D.-Y. CHO and B.-T. ZHANG. *Multi-stage evolutionary algorithms for efficient identification of gene regulatory networks*. in *EvoWorkshops 2006*. 2006: Springer.

[74] S. KIMURA, M. HATAKEYAMA and A. KONAGAYA, "Inference of s-system models of genetic networks from noisy time-series data". Chem-Bio Informatics Journal. **4**(1): p. 1-14, 2004

[75] N. NOMAN and H. IBA, "Inferring gene regulatory networks using differential evolution with local search heuristics". IEEE/ACM Trans Comput Biol Bioinform. **4**(4): p. 634-47, 2007

[76] Y. MAKI, T. UEDA, M. OKAMOTO, N. UEMATSU, Y. INAMURA and Y. EGUCHI, "Inference of genetic network using the expression profile time course data of mouse P19 cells". Genome Informatics. **13**: p. 382-383, 2002

[77] J. S. ALMEIDA and E. O. VOIT, "Neural-network-based parameter estimation in S-system models of biological networks". Genome Inform. **14**: p. 114-23, 2003

[78] E. O. VOIT, S. MARINO and R. LALL, "Challenges for the identification of biological systems from in vivo time series data". In Silico Biology. **5**: p. 0010, 2004

[79] K. Y. TSAI and F. S. WANG, "Evolutionary optimization with data collocation for reverse engineering of biological networks". Bioinformatics. **21**(7): p. 1180-1188, 2005

[80] P. J. SANDS and E. O. VOIT, "Flux-based estimation of parameters in S-systems". Ecol Modeling. **93**: p. 75-88, 1996

[81] G. STEPHANOPOULOS, A. A. ARISTIDOU and J. NIELSEN, "Metabolic Engineering: Principles and Methodologies". 1998, San Diego, CA: Academic Press.

[82] R. HEINRICH and T. A. RAPOPORT, "A linear steady-state treatment of enzymatic chains. General properties, control and effector strength". Eur. J. Biochem. **42**(1): p. 89-95, 1974

[83] G. R. GAVALAS, "Nonlinear Differential Equations of Chemically Reacting Systems". 1968, Berlin: Springer-Verlag.

[84] B. O. PALSSON, "Systems Biology: Properties of Reconstructed Networks". 2006, New York: Cambridge University Press.

[85] M. OKAMOTO. *System analysis of acetone-butanol-ethanol fermentation based on time-sliced metabolic flux analysis*. in *Symposium on Cellular Systems Biology*. 2008. National Chung Cheng University, Taiwan.

[86] N. ISHII, K. NAKAHIGASHI, T. BABA, M. ROBERT, T. SOGA, A. KANAI, T. HIRASAWA, M. NABA, K. HIRAI, A. HOQUE, P. Y. HO, Y. KAKAZU, K. SUGAWARA, S. IGARASHI, S. HARADA, T. MASUDA, N. SUGIYAMA, T. TOGASHI, M. HASEGAWA, Y. TAKAI, K. YUGI, K. ARAKAWA, N. IWATA, Y. TOYA, Y. NAKAYAMA, T. NISHIOKA, K. SHIMIZU, H. MORI and M. TOMITA, "Multiple high-throughput analyses monitor the response of E. coli to perturbations". Science. **316**(5824): p. 593-7, 2007

[87] X. DU, S. J. CALLISTER, N. P. MANES, J. N. ADKINS, R. A. ALEXANDRIDIS, X. ZENG, J. H. ROH, W. E. SMITH, T. J. DONOHUE, S. KAPLAN, R. D. SMITH and M. S. LIPTON, "A computational strategy to analyze label-free temporal bottom-up proteomics data". J Proteome Res. **7**(7): p. 2595-604, 2008

[88] M. VILELA, C. C. BORGES, S. VINGA, A. T. VASCONCELOS, H. SANTOS, E. O. VOIT and J. S. ALMEIDA, "Automated smoother for the numerical decoupling of dynamics models". Bmc Bioinformatics. **8**: p. 305, 2007

[89] H. P. J. BONARIUS, G. SCHMID and J. TRAMPER, "Flux analysis of underdetermined metabolic networks: The quest for the missing constraints". Trends in Biotechnology. **15**(8): p. 308-314, 1997

[90] J. L. REED and B. O. PALSSON, "Thirteen years of building constraint-based in silico models of Escherichia coli". Journal of Bacteriology. **185**(9): p. 2692-2699, 2003

[91] N. ISHII, Y. NAKAYAMA and M. TOMITA, "Distinguishing enzymes using metabolome data for the hybrid dynamic/static method". Theor Biol Med Model. **4**: p. 19, 2007

[92] J. NETER and W. WASSERMAN, "Applied Linear Statistical Models". 1974, Homewood, IL: Richard D. Irwin.

[93] A. RAMOS, A. R. NEVES and H. SANTOS, "Metabolism of lactic acid bacteria studied by nuclear magnetic resonance". Antonie Van Leeuwenhoek. **82**(1-4): p. 249-61, 2002

[94] A. R. NEVES, R. VENTURA, N. MANSOUR, C. SHEARMAN, M. J. GASSON, C. MAYCOCK, A. RAMOS and H. SANTOS, "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. **277**(31): p. 28088-98, 2002

[95] A. R. NEVES, A. RAMOS, C. SHEARMAN, M. J. GASSON, J. S. ALMEIDA and H. SANTOS, "Metabolic characterization of Lactococcus lactis deficient in lactate dehydrogenase using in vivo 13C-NMR". Eur J Biochem. **267**(12): p. 3859-68, 2000

[96] A. R. NEVES, A. RAMOS, H. COSTA, S. VAN, II, J. HUGENHOLTZ, M. KLEEREBEZEM, W. DE VOS and H. SANTOS, "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. **68**(12): p. 6332-42, 2002

[97] A. RAMOS, A. R. NEVES, R. VENTURA, C. MAYCOCK, P. LOPEZ and H. SANTOS, "Effect of pyruvate kinase overproduction on glucose metabolism of Lactococcus lactis". Microbiology. **150**(Pt 4): p. 1103-11, 2004

[98] P. GASPAR, A. R. NEVES, A. RAMOS, M. J. GASSON, C. A. SHEARMAN and H. SANTOS, "Engineering Lactococcus lactis for production of mannitol: high yields from food-grade strains deficient in lactate dehydrogenase and the mannitol transport system". Appl Environ Microbiol. **70**(3): p. 1466-74, 2004

[99] P. H. BERG, E. O. VOIT and R. L. WHITE, "A pharmacodynamic model for the action of the antibiotic imipenem on Pseudomonas aeruginosa populations in vitro". Bull Math Biol. **58**(5): p. 923-38, 1996

[100] I.-C. CHOU and E. O. VOIT, "Recent developments in parameter estimation and structure identification of biochemical and genomic systems". 2009 (submitted)

[101] A. ARKIN and J. ROSS, "Statistical construction of chemical-reaction mechanisms from measured time-series". J. Phys. Chem. **99**(3): p. 970-979, 1995

[102] A. S. TORRALBA, K. YU, P. SHEN, P. J. OEFNER and J. ROSS, "Experimental test of a method for determining causal connectivities of species in reactions". Proc. Natl. Acad. Sci. U S A. **100**(4): p. 1494-1498, 2003

[103] W. VANCE, A. ARKIN and J. ROSS, "Determination of causal connectivities of species in reaction networks". Proc. Natl. Acad. Sci. U S A. **99**(9): p. 5816-5821, 2002

[104] E. SONTAG, A. KIYATKIN and B. N. KHOLODENKO, "Inferring dynamic architecture of cellular networks using time series of gene expression, protein and metabolite data". Bioinformatics. **20**(12): p. 1877-1886, 2004

[105] M. B. EISEN, P. T. SPELLMAN, P. O. BROWN and D. BOTSTEIN, "Cluster analysis and display of genome-wide expression patterns". Proc Natl Acad Sci U S A. **95**(25): p. 14863-8, 1998

[106] K. SACHS, O. PEREZ, D. PE'ER, D. A. LAUFFENBURGER and G. P. NOLAN, "Causal protein-signaling networks derived from multiparameter single-cell data". Science. **308**(5721): p. 523-529, 2005

[107] E. J. CRAMPIN, P. E. MCSHARRY and S. SCHNELL, "Extracting biochemical reaction kinetics from time series data". Lecture Notes in Artificial Intelligence. Vol. 3214. 2004: Springer-Verlag. 329-336.

[108] S. KIKUCHI, D. TOMINAGA, M. ARITA and M. TOMITA, "Pathway finding from given time-courses using genetic algorithm". Genome Informatics. **12**: p. 304-305, 2001

[109] J. SRIVIDHYA, E. J. CRAMPIN, P. E. MCSHARRY and S. SCHNELL, "Reconstructing biochemical pathways from time course data". Proteomics. **7**(6): p. 828-838, 2007

[110] S. EVEN, N. D. LINDLEY and M. COCAIGN-BOUSQUET, "Molecular physiology of sugar catabolism in Lactococcus lactis IL1403". J Bacteriol. **183**(13): p. 3817-24, 2001

[111] M. H. HOEFNAGEL, A. VAN DER BURGT, D. E. MARTENS, J. HUGENHOLTZ and J. L. SNOEP, "Time dependent responses of glycolytic intermediates in a detailed glycolytic model of Lactococcus lactis during glucose run-out experiments". Mol Biol Rep. **29**(1-2): p. 157-61, 2002

[112] B. TEUSINK, J. PASSARGE, C. A. REIJENGA, E. ESGALHADO, C. C. VAN DER WEIJDEN, M. SCHEPPER, M. C. WALSH, B. M. BAKKER, K. VAN DAM, H. V. WESTERHOFF and J. L. SNOEP, "Can yeast glycolysis be understood in terms of in vitro kinetics of the constituent enzymes? Testing biochemistry". Eur J Biochem. **267**(17): p. 5313-29, 2000

[113] W. I. WU, V. M. MCDONOUGH, N. J. T. JR., L. KO, F. A.S., T. R. VLAES, A. H. J. MERRILL and G. M. DARMAN, "Regulation of lipid biosynthesis in *Saccharmomyces cerevisiae* by fumonisin B1". J. Biol. Chem. **270**(22): p. 13171-13178, 1995

[114] K. PESKOV, I. GORYANIN, AND O. DEMIN,, "Kinetic model of phosphofructokinase-1 from Escherichia coli.". Journal of Bioinformatics and Computational Biology,. **6**(4): p. 843-867, 2008

[115] J. HOFMEYR, J. ROHWER, AND J.L. SNOEP,. "Concepts in Computational Systems Biology: Structural Analysis, Kinetics, Control and Regulation of Cellular Systems.". 2007; Available from: http://www.jjj.sun.ac.za/minicourse/).

[116] R. CASTRO, A. R. NEVES, L. L. FONSECA, W. A. POOL, J. KOK, O. P. KUIPERS and H. SANTOS, "Characterization of the individual glucose uptake systems of Lactococcus lactis: mannose-PTS, cellobiose-PTS and the novel GlcU permease". Mol Microbiol. **71**(3): p. 795-806, 2009

[117] A. J. HANEKOM, *Generic kinetic equations for modelling multisubstrate reactions in compuatational systems biology*. 2006: University of Stellenbosch. 114.

[118] F. SHIRAISHI, Y. HATOH and T. IRIE, "An efficient method for calculation of dynamic logarithmic gains in biochemical systems theory". J Theor Biol. **234**(1): p. 79-85, 2005

[119] J. H. SCHWACKE and E. O. VOIT, "Computation and analysis of time-dependent sensitivities in Generalized Mass Action systems". J Theor Biol. **236**(1): p. 21-38, 2005

[120] W. W. CHEN, B. SCHOEBERL, P. J. JASPER, M. NIEPEL, U. B. NIELSEN, D. A. LAUFFENBURGER and P. K. SORGER, "Input-output behavior of ErbB signaling pathways as revealed by a mass action model trained against dynamic data". Mol Syst Biol. **5**: p. 239, 2009

[121] A. MAHDAVI, R. E. DAVEY, P. BHOLA, T. YIN and P. W. ZANDSTRA, "Sensitivity analysis of intracellular signaling pathway kinetics predicts targets for stem cell fate control". PLoS Comput Biol. **3**(7): p. e130, 2007