



FOUNDATIONS OF THEORETICAL APPROACHES IN SYSTEMS BIOLOGY

EDITED BY: Alberto Marin-Sanguino, Julio Vera and Rui Alves

PUBLISHED IN: Frontiers in Genetics, Frontiers in Physiology,
Frontiers in Bioengineering and Biotechnology and
Frontiers in Cell and Developmental Biology



frontiers

Frontiers Copyright Statement

© Copyright 2007-2019 Frontiers Media SA. All rights reserved.

All content included on this site, such as text, graphics, logos, button icons, images, video/audio clips, downloads, data compilations and software, is the property of or is licensed to Frontiers Media SA ("Frontiers") or its licensees and/or subcontractors. The copyright in the text of individual articles is the property of their respective authors, subject to a license granted to Frontiers.

The compilation of articles constituting this e-book, wherever published, as well as the compilation of all other content on this site, is the exclusive property of Frontiers. For the conditions for downloading and copying of e-books from Frontiers' website, please see the Terms for Website Use. If purchasing Frontiers e-books from other websites or sources, the conditions of the website concerned apply.

Images and graphics not forming part of user-contributed materials may not be downloaded or copied without permission.

Individual articles may be downloaded and reproduced in accordance with the principles of the CC-BY licence subject to any copyright or other notices. They may not be re-sold as an e-book.

As author or other contributor you grant a CC-BY licence to others to reproduce your articles, including any graphics and third-party materials supplied by you, in accordance with the Conditions for Website Use and subject to any copyright notices which you include in connection with your articles and materials.

All copyright, and all rights therein, are protected by national and international copyright laws.

The above represents a summary only. For the full conditions see the Conditions for Authors and the Conditions for Website Use.

ISSN 1664-8714

ISBN 978-2-88945-683-3

DOI 10.3389/978-2-88945-683-3

About Frontiers

Frontiers is more than just an open-access publisher of scholarly articles: it is a pioneering approach to the world of academia, radically improving the way scholarly research is managed. The grand vision of Frontiers is a world where all people have an equal opportunity to seek, share and generate knowledge. Frontiers provides immediate and permanent online open access to all its publications, but this alone is not enough to realize our grand goals.

Frontiers Journal Series

The Frontiers Journal Series is a multi-tier and interdisciplinary set of open-access, online journals, promising a paradigm shift from the current review, selection and dissemination processes in academic publishing. All Frontiers journals are driven by researchers for researchers; therefore, they constitute a service to the scholarly community. At the same time, the Frontiers Journal Series operates on a revolutionary invention, the tiered publishing system, initially addressing specific communities of scholars, and gradually climbing up to broader public understanding, thus serving the interests of the lay society, too.

Dedication to Quality

Each Frontiers article is a landmark of the highest quality, thanks to genuinely collaborative interactions between authors and review editors, who include some of the world's best academicians. Research must be certified by peers before entering a stream of knowledge that may eventually reach the public - and shape society; therefore, Frontiers only applies the most rigorous and unbiased reviews.

Frontiers revolutionizes research publishing by freely delivering the most outstanding research, evaluated with no bias from both the academic and social point of view. By applying the most advanced information technologies, Frontiers is catapulting scholarly publishing into a new generation.

What are Frontiers Research Topics?

Frontiers Research Topics are very popular trademarks of the Frontiers Journals Series: they are collections of at least ten articles, all centered on a particular subject. With their unique mix of varied contributions from Original Research to Review Articles, Frontiers Research Topics unify the most influential researchers, the latest key findings and historical advances in a hot research area! Find out more on how to host your own Frontiers Research Topic or contribute to one as an author by contacting the Frontiers Editorial Office: researchtopics@frontiersin.org

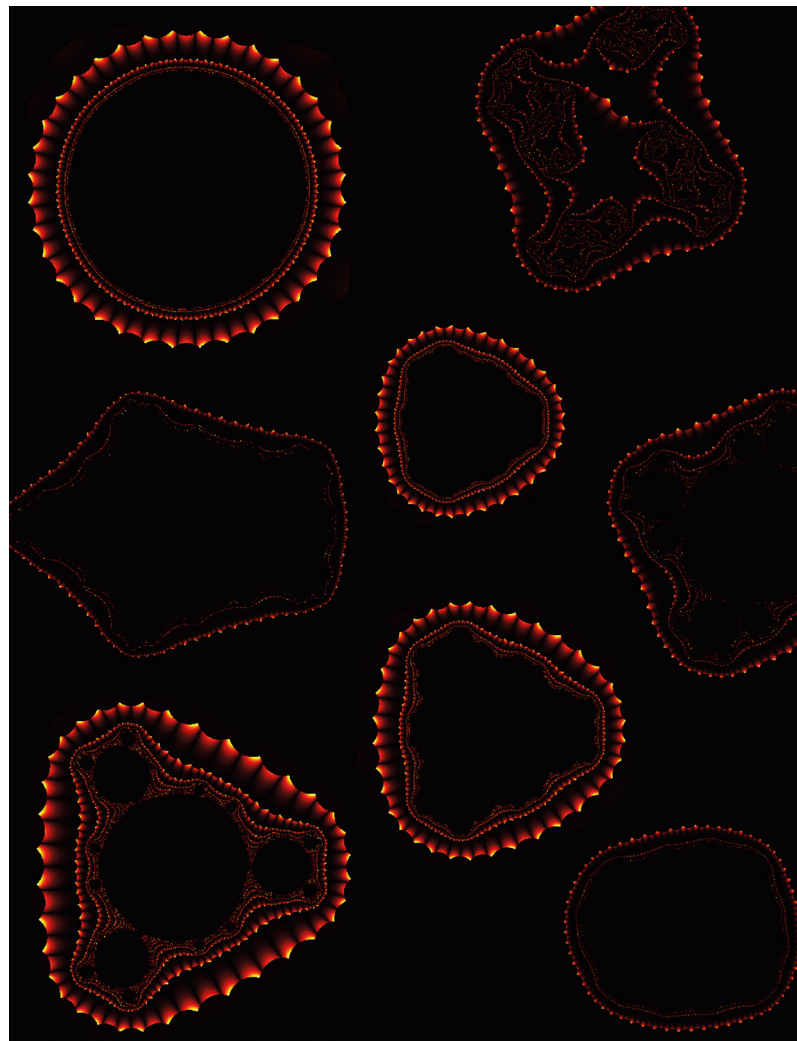
FOUNDATIONS OF THEORETICAL APPROACHES IN SYSTEMS BIOLOGY

Topic Editors:

Alberto Marin-Sanguino, Technische Universität München, Germany

Julio Vera, Universitätsklinikum Erlangen and Friedrich-Alexander-Universität Erlangen-Nürnberg, Germany

Rui Alves, University of Lleida, Spain



Biomorphs generated after the methods developed by C.A. Pickover (<http://www.pickover.com>)

If biology in the 20th century was characterized by an explosion of new experimental technologies, that of the 21st has seen an equally exuberant proliferation of mathematical and computational methods. We are now living through the consolidation of a new paradigm where experimental data goes hand in hand with computational analysis and we must meet the challenge of fusing these two aspects of the new biology into a consistent theoretical framework. Whether systems biology

will survive as a field or be washed away by the tides of future fads will ultimately depend on its success to achieve this type of synthesis. The famous quote attributed to Kurt Lewin comes to mind: "there is nothing more practical than a good theory". This book presents a wide assortment of articles on systems biology in an attempt to capture the variety of current methods in systems biology and show how they can help to find answers to the challenges of modern biology.

Citation: Marin-Sanguino, A., Vera, J., Alves, R., eds. (2019). Foundations of Theoretical Approaches in Systems Biology. Lausanne: Frontiers Media.
doi: 10.3389/978-2-88945-683-3

Table of Contents

| | |
|-----|---|
| 05 | <i>Editorial: Foundations of Theoretical Approaches in Systems Biology</i> |
| | Alberto Marin-Sanguino, Julio Vera and Rui Alves |
| 07 | <i>The (Mathematical) Modeling Process in Biosciences</i> |
| | Nestor V. Torres and Guido Santos |
| 16 | <i>Design Space Toolbox V2: Automated Software Enabling a Novel Phenotype-Centric Modeling Strategy for Natural and Synthetic Biological Systems</i> |
| | Jason G. Lomnitz and Michael A. Savageau |
| 38 | <i>Evolution of Centrality Measurements for the Detection of Essential Proteins in Biological Networks</i> |
| | Mahdi Jalili, Ali Salehzadeh-Yazdi, Shailendra Gupta, Olaf Wolkenhauer, Marjan Yaghmaie, Osbaldo Resendis-Antonio and Kamran Alimoghaddam |
| 42 | <i>Logical Modeling and Dynamical Analysis of Cellular Networks</i> |
| | Wassim Abou-Jaoudé, Pauline Traynard, Pedro T. Monteiro, Julio Saez-Rodriguez, Tomáš Helikar, Denis Thieffry and Claudine Chaouiya |
| 62 | <i>Elementary Vectors and Conformal Sums in Polyhedral Geometry and Their Relevance for Metabolic Pathway Analysis</i> |
| | Stefan Müller and Georg Regensburger |
| 73 | <i>Identification of Metabolic Pathway Systems</i> |
| | Sepideh Dolatshahi and Eberhard O. Voit |
| 90 | <i>A Comparison of Deterministic and Stochastic Modeling Approaches for Biochemical Reaction Systems: On Fixed Points, Means, and Modes</i> |
| | Sayuri K. Hahl and Andreas Kremling |
| 101 | <i>Multiplicity of Mathematical Modeling Strategies to Search for Molecular and Cellular Insights Into Bacteria Lung Infection</i> |
| | Martina Cantone, Guido Santos, Pia Wentker, Xin Lai and Julio Vera |
| 124 | <i>Toward Multiscale Models of Cyanobacterial Growth: A Modular Approach</i> |
| | Stefanie Westermarck and Ralf Steuer |
| 148 | <i>Customized Steady-State Constraints for Parameter Estimation in Non-Linear Ordinary Differential Equation Models</i> |
| | Marcus Rosenblatt, Jens Timmer and Daniel Kaschek |
| 162 | <i>Time Hierarchies and Model Reduction in Canonical Non-linear Models</i> |
| | Hannes Löwe, Andreas Kremling and Alberto Marin-Sanguino |
| 177 | <i>Modeling the Metabolism of Arabidopsis Thaliana: Application of Network Decomposition and Network Reduction in the Context of Petri Nets</i> |
| | Ina Koch, Joachim Nöthen and Enrico Schleiff |
| 199 | <i>Genetic Network Inference Using Hierarchical Structure</i> |
| | Shuhei Kimura, Masato Tokuhisa and Mariko Okada-Hatakeyama |
| 211 | <i>Book Review: Python Programming for Biology</i> |
| | Alberto Marin-Sanguino |
| 213 | <i>Book Review: Computing for Biologists: Python Programming and Principles</i> |
| | Alberto Marin-Sanguino |



Editorial: Foundations of Theoretical Approaches in Systems Biology

Alberto Marin-Sanguino^{1†}, Julio Vera^{2*†} and Rui Alves^{3†}

¹ Specialty Division for Systems Biotechnology, Technische Universität München, Garching bei München, Germany,

² Laboratory of Systems Tumor Immunology, Department of Dermatology, Universitätsklinikum Erlangen and Friedrich-Alexander-Universität Erlangen-Nürnberg, Erlangen, Germany, ³ Departament de Ciències Mèdiques Bàsiques, University of Lleida, Lleida, Spain

Keywords: systems biology, network biology, mathematical modeling, computational modeling, systems medicine, biotechnology

Editorial on the Research Topic

Foundations of Theoretical Approaches in Systems Biology

The importance of systemic approaches in understanding biology was recognized as early as in the nineteenth century (Bernard and Dagonet, 2013). Around the 1920s and over the next few decades Briggs and Haldane (1925); von Bertalanffy (1962) and others Savageau (1969); Michaelis and Menten (2013) showed that such systemic views were both scientific and necessary in the biological sciences. Still, the only technology that could accurately perform biological studies integrating a large number of molecular components was mathematical modeling. This limitation remained in place until the late 1990s, making these studies hard to validate experimentally.

This pre-history of Systems Biology would end when full genome sequencing and the high throughput methods that would follow flooded every biological discipline with more data than could be analyzed. As a consequence, many discarded the usefulness of mathematical modeling under the assumption that there is no need to simulate what can be measured. Over time this view was understood as simplistic, and it became clear that mathematical and statistical modeling is essential to distill the sheer amount of molecular data available into “general biological laws” that explain how molecular components come together and form biological systems. We are leaving an era where large scale measurements of all molecular components in a cell dominated the field and entering a new wave of methodological development to integrate all those measurements into meaningful mathematical descriptions.

This integration needs to be multilevel. We need accurate methods that use experimental and qualitative information to perform whole-genome network reconstruction at the metabolic, signaling and the gene regulation level. We need general techniques that automatically derive and analyze mathematical models of such reconstructed networks. This Frontiers research topic, “Foundations of Theoretical Approaches in Systems Biology,” aims at paving the way to investigate if this set of approaches is mature enough to coalesce into a coherent body of knowledge.

In line with this, Torres and Santos introductory paper outlines the traditional modeling process as three-stage framework. In the first stage the biological system is framed as a conceptual model. In the second stage, the model is represented using a formal mathematical description. In the final stage the mathematical description is parameterized and studied through analytical and simulation methods to understand the dynamic behavior and regulation of the system. Lomnitz and Savageau recognize the limitations implicit to that classical approach. They describe a method in which all possible qualitatively different types of dynamical behavior, or phenotypes, of the system can be mapped from the conceptual representation and identify the parameter ranges that make each phenotype realizable. They also contribute a toolbox that enables modelers to try that method.

Other contributions to the topic describe and analyze the diversity of modeling being used and emphasize some of the commonalities and differences among them. At the level of network reconstruction, where little quantitative information is available, network centrality measures

OPEN ACCESS

Edited and reviewed by:

Raina Robeva,
Sweet Briar College, United States

*Correspondence:

Julio Vera
julio.vera-gonzalez@uk-erlangen.de

[†]These authors have contributed
equally to this work

Specialty section:

This article was submitted to
Systems Biology,
a section of the journal
Frontiers in Genetics

Received: 13 June 2018

Accepted: 12 July 2018

Published: 15 August 2018

Citation:

Marin-Sanguino A, Vera J and Alves R
(2018) Editorial: Foundations of
Theoretical Approaches in Systems
Biology. *Front. Genet.* 9:290.
doi: 10.3389/fgene.2018.00290

determined using graph theoretical approaches can help in identifying the key elements in the network, as is reviewed by Jalili et al.. As the causal structure of the network becomes clearer, logic modeling can offer testable dynamical and regulatory insights about the way in which, for example, signaling and gene regulation networks work (Khan et al., 2017). Abou-Jaoudé et al. review and discuss the potential of this type of modeling to reconstruct and analyze large, intricate biochemical networks.

Moving to models that describe biological systems using linear mathematics and steady state approximations, Müller and Regensburger explores the concept of elementary flux modes, a defining set for every possible flux distribution in a biochemical network. They use combinatorial mathematics and polyhedral geometry (Rockafellar, 1969) to propose alternative ways to search for flux modes in metabolic network analysis. Dolatshahi and Voit explore and discuss strategies for model parameters estimation that extend the use of *dynamic flux estimation* method for the analysis of metabolic time series data to general, slightly underdetermined metabolic networks. This method establishes a bridge between constraint-based models, which can be formulated with minimal information, and kinetic models that can be used to analyze transient data.

Hahl and Kremling examine the parallels and discrepancies between deterministic (ordinary differential equations) and stochastic approaches (chemical master equation) of molecular systems, discussing when to choose one or the other.

Overall, choosing a modeling framework is a trade-off that should consider the question being addressed as well as the data that is available to inform model creation. Models for bacterial lung infection (Cantone et al.) and cyanobacteria (Westermarck and Steuer) are used to illustrate the advantages and disadvantages of alternative approaches, and to point out ways in which those approaches can be combined to create multi-level models.

Another important issue in mathematical modeling is that of model reduction. This is the process of identifying simpler but accurate enough versions of a larger model. Classical approaches to model reduction can be found in the field of enzyme kinetics. This field combines graph theoretical approaches with considerations about the differences between the characteristic

time scale of individual chemical reactions or between the concentrations of the various species in a network to derive single equations that describe the dynamic behavior of fairly complex networks. Rosenblatt and coworkers (Rosenblatt et al.) present a graph-theoretical algorithm for deriving steady-state expressions by stepwise removal of cyclic dependencies between the network model variables. In parallel Löwe et al. and Koch et al. provide examples that illustrate the importance of choosing the appropriate mathematical formalism and how that formalism can be used to develop efficient approaches to model reduction.

Coming full circle, Kimura et al. illustrate that dynamic mathematical models can also be used for inferring network structure and refining the initial conceptual model on which the mathematical model is based.

Together, the collection of papers under the research topic “Foundations of Theoretical Approaches in Systems Biology” shows how theoreticians are exploring many different avenues to interpret experimental data and distill them into “biological laws.” In addition, this topic contributes to understand where those approaches overlap and where they complement one another. Only through such an effort can we avoid fragmentation and minimize duplication of efforts, and thus contribute to the consolidation of Systems Biology as a field of knowledge rather than an assortment of techniques.

AUTHOR CONTRIBUTIONS

All authors listed have made a substantial, direct and intellectual contribution to the work, and approved it for publication.

FUNDING

AM-S was funded by the German Ministry of Education and Research (BMBF) projects OpHeLiA (0316197) and HOBBIT (031B0363A). RA funded by Generalitat de Catalunya Consolidated Group SGR133 (2017). JV was funded by the German Ministry of Education and Research (BMBF) projects e:Med-CAPSyS (01ZX1604F and 01ZX1304F) and e:Bio-MeEVIR (031L0073A).

REFERENCES

- Bernard, C., and Dagonet, F. (2013). *Introduction à l'étude de la Médecine Expérimentale*. Flammarion.
- Briggs, G. E., and Haldane, J. B. (1925). A note on the kinetics of enzyme action. *Biochem. J.* 19, 338–339.
- Khan, F. M., Marquardt, S., Gupta, S. K., Knoll, S., Schmitz, U., Spitschak, A., et al. (2017). Unraveling a tumor type-specific regulatory core underlying E2F1-mediated epithelial-mesenchymal transition to predict receptor protein signatures. *Nat. Commun.* 8:198. doi: 10.1038/s41467-017-00268-2
- Michaelis, L., and Menten, M. M. (2013). The kinetics of invertin action. *FEBS Lett.* 587, 2712–2720. doi: 10.1016/j.febslet.2013.07.015
- Rockafellar, R. T. (1969). “The elementary vectors of a subspace of R^N ,” in *Combinatorial Mathematics and Its Applications (Proceedings of Conference, University of North Carolina, Chapel Hill, N.C., 1967)* (Chapel Hill, NC: University of North Carolina Press), 104–127.
- Savageau, M. A. (1969). Biochemical systems analysis. I. Some mathematical properties of the rate law for the component enzymatic reactions. *J. Theor. Biol.* 25, 365–369.
- von Bertalanffy, L. (1962). *Modern Theories of Development an Introduction To Theoretical Biology, 1st Edn.* New York, NY: Harper. Available online at: <https://www.amazon.co.uk/Theories-Development-Introduction-Theoretical-Biology/dp/B0007E65IK> (Accessed June 11, 2018).

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2018 Marin-Sanguino, Vera and Alves. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



The (Mathematical) Modeling Process in Biosciences

Néstor V. Torres^{1,2*} and Guido Santos^{1,2}

¹ Systems Biology and Mathematical Modelling Group, Departamento de Bioquímica, Microbiología, Biología Celular y Genética, Sección de Biología de la Facultad de Ciencias, Universidad de La Laguna, San Cristóbal de La Laguna, Spain,
² Instituto de Tecnología Biomédica, CIBICAN, San Cristóbal de La Laguna, Spain

In this communication, we introduce a general framework and discussion on the role of models and the modeling process in the field of biosciences. The objective is to sum up the common procedures during the formalization and analysis of a biological problem from the perspective of Systems Biology, which approaches the study of biological systems as a whole. We begin by presenting the definitions of (biological) system and model. Particular attention is given to the meaning of mathematical model within the context of biology. Then, we present the process of modeling and analysis of biological systems. Three stages are described in detail: conceptualization of the biological system into a model, mathematical formalization of the previous conceptual model and optimization and system management derived from the analysis of the mathematical model. All along this work the main features and shortcomings of the process are analyzed and a set of rules that could help in the task of modeling any biological system are presented. Special regard is given to the formative requirements and the interdisciplinary nature of this approach. We conclude with some general considerations on the challenges that modeling is posing to current biology.

OPEN ACCESS

Edited by:

Rui Alves,
Universitat de Lleida, Spain

Reviewed by:

Miguel Ángel Medina,
University of Málaga, Spain
Ester Vilaprinyo,
University of Lleida – Institut
de Recerca Biomèdica de Lleida,
Spain

*Correspondence:

Néstor V. Torres
ntorres@ull.edu.es

Specialty section:

This article was submitted to
Systems Biology,
a section of the journal
Frontiers in Genetics

Received: 23 September 2015

Accepted: 07 December 2015

Published: 22 December 2015

Citation:

Torres NV and Santos G (2015)
The (Mathematical) Modeling Process
in Biosciences. *Front. Genet.* 6:354.
doi: 10.3389/fgene.2015.00354

Keywords: biosciences, biological system, model, mathematical model, systems biology

INTRODUCTION

A theory has only the alternative of being right or wrong. A model has a third possibility: it may be right, but irrelevant.

Manfred Eigen. The Origins of Biological Information.

There are many definitions of science (Popper, 1935; Kuhn, 1962, 1965; Lakatos, 1970), but all of them refer to a body of knowledge obtained through a particular method based on the observation of the physical world, linked to systematically structured reasoning, strategies by which general principles and laws are deduced. That particular method is the “Scientific Method”, defined by the Oxford English Dictionary as “...the procedure... consisting in systematic observation, measurement, and experiment, and the formulation, testing, and modification of hypotheses.” In the above statements there are two core ideas which are relevant here and that derive directly from what science is: the first one is that any scientific activity requires measurements and thus, quantification of real magnitudes. The second is that any scientific activity makes sense only if it allows us to gain “knowledge”; that is understanding, predicting and control. In science these goals are achieved through the building of models and theories. Both serve, with different degrees of generality, to explain the observed facts and predict with high probability the evolution and behavior of natural systems.

Biological Systems and Models

Before describing the modeling process, it is advisable to clarify the meaning of two key concepts, “biological system” and “model” that we assume are inextricably linked.

Any biological system is composed of a set of elements, physical objects, usually numerous and diverse, that influence each other (i.e., they interact) and that are physically and functionally separated from their environment. The physical separation is a frontier, which can be real (e.g., a membrane) or imaginary, which is permeable to matter, and energy (i.e., an open system). The functional separation is a consequence of the fact that biological systems are far from thermodynamic equilibrium, in contrast with the environment. The interchange of matter and energy with the environment is indeed a necessary requisite to sustain the chemical–physical processes that occur far from equilibrium. Thus defined, a living system involves a reference to the environment in which it is located and with which it interacts. It is worth noting here that when we focus solely on the elements, disregarding the interactions between them and with the environment, the system disappears, because a set of entities devoid of interaction is a mere aggregation of elements. This is the essence of “system”, a holistic approach to research as opposite to a reductionist view.

For our purposes here, a model is a conceptual or mathematical representation of a system that serves to understand and quantify it. The difference between conceptual and mathematical resides only on the way the representation is formulated. A model is always a simplified representation of the reference system, which the scientist wishes to understand and quantify. It ultimately serves as a means of systematizing the available knowledge and understanding of a given phenomenon and the facts concerning it.

A first step in any model-building attempt is the simple verbalization of statements about the biological system. Soon this phase leads to a more productive one, where observations and hypothesis transform the observations and data into an organized core, the so-called “conceptual” model. Conceptual models constitute, thus, a first level of qualitative integration of the information on the system under scrutiny. Conceptual models are so ingrained in our everyday life that we usually do not make a distinction between models and the real thing. Very often, they come as diagrams, words or physical structures, which deal with either the structure and/or the function of the real system. The causal diagrams are examples of suitable tools that help in dealing with the conceptual models (Voit, 1992; Minegishi and Thiel, 2000; Allender et al., 2015).

A key feature of the conceptual models is that they only make a qualitative description of the real system. Examples of such conceptual models in biology range from the typical plant or animal cell diagram (one that integrates many observations of multiple types of cells obtained through a great variety of techniques) to the models about enzyme action and metabolic pathways. The enzyme action model describes how the substrate attaches to the active site of the enzyme, and how the enzyme structure changes in different molecular environments. Another ubiquitous conceptual model is that of metabolic pathways;

they represent the coordinated and sequential activities and regulatory features of many enzymes. The main value of the conceptual models is that, as the result of the (tough) complex process involved in its development, it allows the integration of disperse information obtained from different sources. However, their origin renders them imprecise, and conceptual models can be interpreted differently by different people.

A further refinement in the process of system understanding is given by the translation of the conceptual model into a form subject to a quantitative description, evaluation and validation. This form is the mathematical model. A mathematical model is the formalized description of the system derived from a previous conceptual model. Mathematical models may be very diverse in nature. Dynamical models consider changes in the elements with time, and can be categorized into deterministic and stochastic. In the deterministic ones, the velocities only depend on the concentration of the elements and the parameters of the model. The opposite are the stochastic ones, in which the velocities also depend on the random noise of the system, due to the uncertainty present in systems containing statistically non-abundant elements. On the other hand, static models try to understand the structure of the interconnection of the elements, which remains constant during time under specific conditions (Voit, 2012).

The mathematical models not only help us to understand the system, but also are instrumental to yield insight into the complex processes involved in biological systems by extracting the essential meaning of the hypotheses (Wimsatt, 1987; Bedau, 1999; Schank, 2008) and allows to study the effects of changes in its components and/or environmental conditions on the system's behavior; that is, they allow the control and optimization of the system.

Mathematical Models in Biology

The usefulness of mathematical models in physics and technology is well documented; in fact they can be traced back to the very origins of physics. Since the days of Galileo, Kepler and Newton scientists have striven to develop their models by means of mathematical formalism. What we want to present and develop here is the tenet that modeling in general, but specifically mathematical modeling, particularly in biology –as well as in science in general– is the only way to attain such quantitative understanding and control. Mathematical modeling should thus be an essential and inseparable part of any scientific endeavor in the realm of XXI century bioscience.

It has been claimed that the maturity of a scientific field correlates positively with how often mathematical models are developed and used to understand and control the real system (Weidlich, 2003; Medio, 2006; Brauer and Castillo-Chavez, 2010; Gunawardena, 2011). In this regard, it has not been until recently that dynamic mathematical models in biology have become a common feature. Besides the well-known cases of the Michaelis–Menten model to describe the dynamics of the enzyme-catalyzed reactions (Michaelis and Menten, 1913) and its subsequent development for the case of allosteric enzymes (Monod et al., 1965), the Hodgkin–Huxley model of the action potentials in

neurons (Hodkin and Huxley, 1952), the Lotka–Volterra model about the interaction of species (Lotka, 1920; Volterra, 1926) and the epidemiological models of epidemics (Ross, 1915; MacDonald et al., 1968), the emergence and widespread recognition of the role and importance of mathematical models in biology is a recent phenomenon.

It is easy to understand why only until very late in scientific research mathematical modeling of biological systems has been put in use. Biological systems, by their nature, are refractory to precise quantitative and mathematical description. They are composed by many elements closely interconnected by processes and interactions that take place at different levels of organization (molecular, cellular, in tissue, whole animals and ecological). At the same time, these processes occur in an open system as a result of the existence of multiple gradients far from the thermodynamic equilibrium, which in the end produce very complicated non-linear dynamics between the elements of the system (Prigogine, 1961). This situation has impaired the quantitative and dynamic approach to the understanding of biological systems through the use of mathematical models.

However, two technological advancements that have made feasible the construction and resolution of mathematical models for biological systems have been developed in the last decades. There is a general accessibility and almost universal ubiquity of the computational power required for the management of information and the calculation of large systems. On the other hand, the development of the high throughput techniques and the emergence of the “omics” sciences (genomics, transcriptomics, proteomics, signalomics, and metabolomics) have generated a great deal of dynamic information on the structure and behavior of the biological systems. This information has become easier and cheaper to acquire, process and store than ever before.

All the above have been instrumental to the arrival of Systems Biology, as the XXI century approach to the quantitative and interdisciplinary study of the complex interactions and the collective behavior of a cell, an organism or an ecosystem. The distinctive feature of Systems Biology is the concern with the organization and biological function. This approach goes beyond the classical reductionist approach, where the researcher seeks to understand the systems by breaking them down into their constituent elements and analyzing them separately or, in a novel version of the old paradigm facilitated by the high throughput techniques, by collecting every piece of accessible information. In the Systems Biology approach, research is focussed not on the parts considered individually, but on the relationships that exist between the structural components of biological systems and their function, and on the characteristics of the interactions that occur between different sub-systems. This method allows the detection of emerging higher levels of structural and functional organization. In contrast with the reductionist approach, Systems Biology deals with the reconstructive and integrative task upon the available biological information. And it is here where models and modeling becomes a central tenet in Systems Biology.

In the following section we will develop a general framework where the role of models and the modeling process within the scientific activity in biosciences is highlighted. Also, a set of rules

that help the modeling activity is presented together with some general considerations on the challenges that modeling currently poses.

A MODEL OF THE MODELING PROCESS IN BIOSCIENCES

The purpose of models is not to fit the data but to sharpen the questions.

Samuel Karlin

The **Figure 1** summarizes the set of activities and elements involved in the development of models, as organized following the Scientific Method.

I. Conceptualization

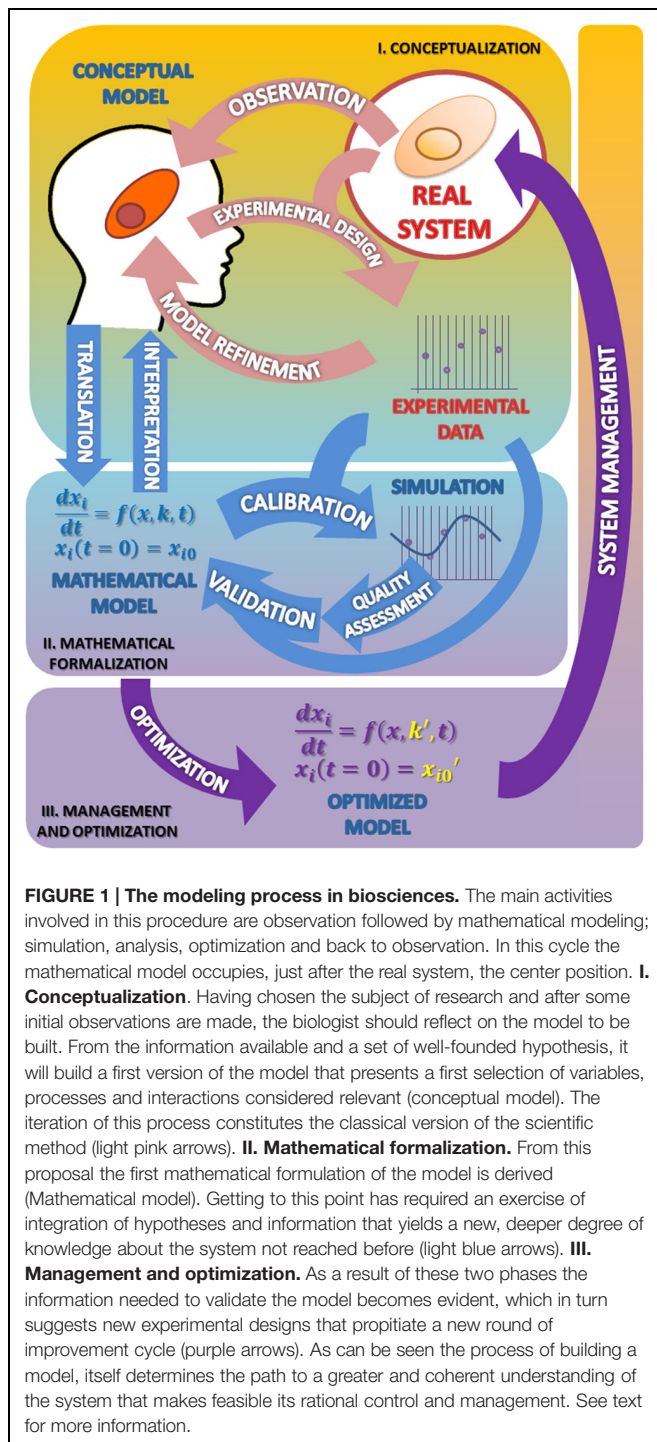
The first stage of the scientific modeling process is the conceptualization phase. In any research process all activities are organized around the Real System, which is the compulsory, continuous reference in the whole process. This central position is represented in **Figure 1** as a circle.

The first step in the conceptualization stage is to formulate, from the very first observations of the phenomenon (Observation; see **Figure 1**), generally made in an unsystematic form, an explanatory hypothesis of it: the first version of the conceptual model. This is a critical task where it is necessary to coordinate, to contrast and discuss many issues with the aim of making the best decisions. Some of the questions that should be addressed at this stage are: what aspects of the real system should be incorporated into the model? What features should/can be ignored? Or, what hypotheses can support the observations/information rendered by the system?

Given that any model is an instrument designed for a purpose, the very first question that should be posed at this stage is: what is the model for? That is, the objective of the model. No model makes sense or is justified for its own sake. Thus, what first defines a model is the specific question that it is going to answer.

Trying to develop a model to explain all aspects of a biological phenomenon will be practically impossible, a very complex and highly unmanageable task. However, a model with a limited purpose will be feasible, and easier to be analyzed and managed. At this stage of modeling, our thinking process uses the categories of space, time, substance (namely, material components, and elements), quality, quantity, and relationship. These categories help us to bring order to the perceived complexity of the real world. Nevertheless, this act of classification and identification differ considerably from one scientific discipline to another.

The meaning and significance of the modeling process is rooted in the core of the scientific process: from the observation of some part of the biological world some questions arise, the model being the tool that eventually would serve to provide an answer. As can be seen, any modeling exercise forces, from the very beginning, to define and make explicit the focus of our research and to keep, all along the way, our attention on the main objective.



The conceptualization stage is where modeling becomes very often an art, a subjective task. The choice of the essential attributes of the real system and the omission of irrelevant ones requires a selective perception that you cannot specify through an algorithm. There is some dosage of freedom and arbitrariness at this stage since different researchers equally well informed can define different models. As we are educated in a specific biological scientific discipline, we are trained to

observe the real world in the light of a certain conceptual framework.

In some instances, the discussion of contrasting opinions addressed to demarcate the border between the system and its environment, or to discriminate between different possible scenarios or to evaluate the importance of the experimental error associated with the observed values, leads to different versions of the model. Based on the final selection of hypotheses, the next step is to carry out experiments (Experimental design; see **Figure 1**) devised to obtain experimental data to test the chosen hypothesis. From the analysis of the experimental results, the hypothesis can be reformulated or discarded (Model refinement; see **Figure 1**), thereby initiating a virtuous cycle (pink arrows) that leads to an improved conceptual model. Eventually, this refined model version is expected to answer, though qualitatively, the questions initially raised. At this stage, the need to change the initial hypothesis, far from being a failure, should be understood as progress toward a better understanding of the behavior of the system. This allows to rule out some proposals, which will be replaced by new ones that might be more effective in the building process of the conceptual model.

The above sequence illustrates the fact that observation and science are not the same thing. The aim of the scientific method is not to describe but to explain the observed, to understand and interpret the observations. It is here where the collaboration between the modeling part and the field experts becomes essential. And it is at this stage where interdisciplinarity occurs. The best version of the modeling task results when it is a team effort, where the competences and expertise of different specialists blend. Those with the best knowledge on the particular subject should be able to communicate with the modeler. They must be able to understand each other; the expert presenting the whole picture and selecting from it the elements, interactions, processes, and values that are deemed relevant in the light of the model's objective. At this point the modeler should translate this selection into a conceptual representation that usually takes the form of a mechanistic picture where the elements and their relations are represented. To be useful, this picture should be explicit enough to be translated into a series of elementary steps representing the individual mechanisms. The modeler here is instrumental in defining which are the magnitudes considered as variables and which are not; this is a critical distinction that determines to a great extent the model's output.

The development of the modeling approach has at this point one of its great challenges, because it requires that the different specialists share a common language. There is a need, on the side of the modeler, to become acquainted with the features and nuances of the system under scrutiny, and to speak in terms easy to understand by the non-modeler party. On the other side, the specialist should adopt an integrative way of thinking and be able to make explicit his knowledge and express it in the most precise terms.

More often than not, it is necessary to repeat the conceptualization stage of discussion and analysis several times, before the proposed model becomes able to respond

successfully to all the objections that could be raised by the experts who come in contact with the model. Once you have reached an acceptable version you will be able to consider the next stage: the mathematical formalization.

At this phase of the model building process it could happen that the modeler may be tempted by the challenge of building a wholly comprehensive model system, that is, one that takes into account, if not all, most of the characteristics of the real system. Besides the misunderstanding of the modeling process that this shows, this attitude has additional costs, because if two models serve to give the desired answers, the simpler one is better. A modeler intending to include all variables and parameters described would also be faced with the task of analyzing the influence of all the parameters on all the variables. This in turn would require an additional, usually non-negligible effort for its interpretation, making the model more difficult to understand. In modeling, more and harder is not necessarily better. In fact it sometimes happens that the largest and most complicated model may be the poorest in attaining its objectives or expressing necessary or meaningful details of the reality. A nice illustration of this point is the very simple model of the signaling pathway of NF- κ B, in which with only three elements it is reproduced the main dynamical behavior of the original system (Krishna et al., 2006). In other words, we should try to make the complex as uncomplicated as possible. Despite this, the discussion of its results can enrich the conceptual model building when considering the traits and characteristics that were not initially included.

Related with this is the fact that the developments of the conceptual model force the analysis and the systematic review of available knowledge about the system and its behavior. As a result of this exercise of verbalization of the knowledge -often unconscious- that experts have about the system, a new light is shed on the phenomenon, which very often contributes to a better understanding of the system.

It may happen that some gaps of information about interactions or relevant parts that had hitherto gone unnoticed become evident. This usually suggests new avenues of exploration and ultimately contributes to a better understanding of the observed reality. Also, the discussions on the variables or the processes involved help to change previous assumptions or facilitates a new view and understanding of some facts that previously remained without an explanation. As an example, Cheong et al. (2008) review the contributions of mathematical modeling on the understanding of the NF- κ B pathway. It is also very common to become aware of contradictions in the understanding of biological mechanisms. Most of the knowledge or information about an issue may pass through several authors undisputed, but when all this is mathematically formalized, problems to join all in a single framework emerge. Mathematical thinking forces to reconsider every piece of knowledge.

Finally, there is a modeling principle that should be commented here: "If the hypotheses of the model are erroneous, the conclusions raised from it will be wrong too." As obvious as it may be, this principle is not less important. This principle should be taken into account all along the model-building process, particularly in the mathematical formalization that follows,

because the resulting model should be faithful to the proposed hypothesis.

II. The Mathematical Formalization

Mathematical Translation

The first question to be addressed in this new phase is about which mathematical formalism is best suited to represent the system (Translation; see **Figure 1**). There are many formal modeling approaches, based on differential equations, Bayesian equations, stochastic systems, agent-based modeling, etc. (for a review, see ElKalaawy and Wassal, 2015). Each of these has unique strengths and limitations. The choice heavily depends on the nature of the model. It often happens that a research group ends up enslaved by the modeling techniques which it dominates or prefers. For example, a team with experience in modeling using differential equations may tend to approach every problem from the standpoint of this technique, when in fact not all biological problems are deterministic. It is natural to preferentially use the methods that are best known and previously proven fruitful. But the ideal attitude is to adapt the specific modeling technique to the nature of the problem.

The task of developing a model is a process of approximation due to the simplifications that must be introduced. These simplifications should make sense in terms of the physical-chemical processes being studied, but must also be valid from a mathematical point of view. The general approach to the mathematical formulation usually involves the definition of the key variables and the expression of their functional relationship with the other variables of the system. Equations are then derived establishing the actual mathematical relationship among the variables. This derivation can be done empirically (data-driven), through the use of statistical methods (curve fitting) analytically or numerically, or by deriving the equations from theoretical considerations (model-driven). A classic example of model-driven is given by Michaelis and Menten (1913) kinetics. Other common techniques of data-driven modeling are shown in Janes and Yaffe (2006). In the model we should make clear the differences among the variables (concentrations of biochemical compounds of the investigated network: metabolites, proteins, messenger RNAs, etc.) and the parameters. Variables can be dependent, being the elements which vary over time according to the state of the system (also called states); and independent, being the ones that can be controlled during the experiment (light, pH, etc.). The parameters set internal and external constraints on the system. The specific numerical values for the parameters are determined using prior biological knowledge, such as information about the basal steady states of the system (Voit, 2000), or experimental data from dynamical perturbations (Vera et al., 2007, 2008). Usually the models integrate kinetic data and other available information about the elements of the process, as well as fluxes obtained from experimental observations.

It often happens that an existing model is used to describe another system. This strategy, although tempting, should be used with caution. Each new system should be studied in their specific conditions of environment and structure. It is also necessary to consider that a model not only depends on the system that it

represents and the techniques used for its construction, but also from the motivations and objectives of their creators. Therefore one must always beware from attributing the motivations and objectives of others to our own model.

The process of developing a mathematical formulation of the conceptual model forces the investigator to describe the system in simple terms. At this stage the research team must take into account details about the system which might otherwise go unnoticed, which contribute to the improvement of the model. Also, a healthy consequence of the formalization process is that the explanations of the initial, sometimes unexamined assumptions reveal processes and features that remained unrecognized under the less precise conceptual formulations.

The interpretation and understanding of the system has an additional resource in the mathematical expression of it (see **Figure 1**). The set of equations of the mathematical model is likely to be discussed with the plethora of techniques and mathematical tools that allow the description and analysis of the complex interrelated processes that occur in the real system; these techniques can help to elucidate the structure, properties, and dynamic behavior of the system. These analyses can reveal details about the behavior of a model such as the occurrence of oscillations or other complex behaviors that are often the motivating force for studying these systems.

Parameter Estimation and Quality Assessment

Once the conceptual model has been translated to its mathematical form, the model should be provided with the values of its parameters. Parameter estimation or model calibration is a recurrent issue in the model building process; it deals with the finding of the numerical values which characterize the mathematical representation of a given system from experimental data (Park et al., 1997). A key feature of these experimental measurements is that they must come from variables representing their main features both at a given particular time, as well as along its evolution over time (Polisety and Voit, 2006; van Riel, 2006; Ashyraliyev et al., 2008; Banga, 2008). In addition, the quality of the model should be tested through some numerical quality assessments. The quality assessment of the model includes the evaluation of aspects such as the stability of steady states, a prerequisite for any model describing actual biological systems; and the robustness of the model, a test to evaluate whether the model is able to tolerate small structural changes (Savageau, 1971; Hsiung et al., 2008) and the dynamic features that characterize the transient responses to temporary perturbations or permanent alterations (Wu et al., 2008). These analyses often pinpoint problems of consistency and reliability of the mathematical representation (Okamoto and Savageau, 1984, 1986; Ni and Savageau, 1996a,b); this constitutes by itself a further cycle of model refinement (**Figure 1**, light blue cycle). These changes result in improvements of the initial conceptual model. The conceptual model so improved will in turn suggest further experimentation leading to a new refined version that is enriched from the formalization phase.

At all instances it should be borne in mind that both the parameters and the structure of real systems change over time. Therefore, a given model, which can be satisfactory at one time

or certain conditions, may lose its effectiveness at another time or in other conditions. But the equations by themselves do not contribute much to the understanding of the system. It is necessary to solve the equations for some representative values of the parameters. Accordingly, the model is submitted to the simulation and validation processes.

Simulation and Prediction

The mathematical model should be programmed in the computer. The computer program is the translation of the mathematical model to another language useful for computing purposes. There are many computer languages and platforms to deal with this task; advances in computer numerical analysis have made the solution of complicated systems fast and accurate. It is at this point where computation becomes critical, since the equations describing biological processes nearly always involve control and regulatory mechanisms that have non-linear components. In contrast with linear differential equations that often can be solved analytically, non-linearities make it generally impossible. But through the use of numerical methods implemented on computers we can obtain good estimates and model predicted data.

Model Validation

Validation stands here as the correspondence between the real system and the mathematical model. A model can be considered good and useful only if its predictions in a given scenario agree with the experimental observations made on the actual system setting. As it is shown in the **Figure 1**, we can accept the model as a plausible representation of the system under scrutiny only when the comparison of the predicted outputs with the real ones yields similar results (and when this occurs in a significant number of situations).

The validation process can only be based on comparative observations of the output values and trajectories of the model and the real system, under certain given experimental conditions. As it is shown in the **Figure 1**, for validation purposes, a first cycle of calibration and quality assessment is needed, and then a second one, with new experimental data from a different condition. As a result, the model might require some modification in order to minimize the observed discrepancies.

There are several ways to perform the validation process. One is to compare the evolution of the variables from some, other initial conditions; with data obtained by different, other research groups in similar systems. Another way is to use all available data in some given conditions, not for the development of our own model, but to use these data for the comparison with our model's predictions instead (Santos and Torres, 2013). In some cases, a useful technique is to vary some model's parameters within the range of biologically plausible values, and observe how the system responds in relation to its actual behavior (Segre et al., 2002). Finally, a technique that can be used in some instances involves subjecting the model to extreme conditions, deliberately looking for their failures. The logic behind this is that, if a model represents the system well in extreme conditions, so it will under normal conditions. In any instance the observed discrepancies indicate errors in the assumptions used in the building of the

mathematical and/or the conceptual model. The discrepancies may be large enough as to require the revision and change of the hypothesis of the conceptual model, or the introduction of only slight modifications in the parameters of the mathematical version.

It should be recalled that the quality of a model depends directly on the quality of the information on which it is based. This statement is just the translation to the modeling context of the classical motto: “garbage in, garbage out”. A mathematical model logically processes the information with which it has been built; it cannot recognize or correct errors in the definitions or the input information. The model predictions are the result of the assumptions used, hence the extreme importance of caring for the conceptualization phase and the quality of the initial information.

Very often the most important results of this phase are negative: a well-crafted model would serve to discard a particular mechanism as the explanation for experimental observations. After sufficient validation, we will eventually arrive to a mathematical version of the model that can be used to perform experiments in much the same manner as in the real system. A model is considered valid in this respect when the decisions made using the mathematical model are “similar” to those that would have been made by physically experimenting with the real system.

This model version and its computer counterpart allow testing conditions that may be difficult to attain in the laboratory, or that have not yet been examined in actual experiments. Each numerical solution of the model provides a simulation of a real or potential experiment carried out on the Real System. As an example of mathematical simulations which were useful to understand the dynamics of the cell membrane, a biological process elusive under laboratory experiments see Marrink and Tieleman (2013).

In this phase, starting from a first version of the mathematical model we come to an improved, validated version, through a new virtuous cycle (light blue arrows) that sum up to the first one (light pink arrows). Repeated excursion through this research loop can result in further improvements in both the mathematical and the conceptual model that provides an unbiased test of the hypothesis involved in the conceptual model. This type of feedback loops, which are an essential part of the process of developing a model (and indeed of the scientific method), must, however, stop at some point. The validation phase often leads to a situation in which a slight increase in the trust of the model requires a huge effort. In these cases, it is advisable to stop the process at this point.

Model Refinement and Interpretation

Once we have reached a sound mathematical version of the real system we can advance in its interpretation and understanding. At this stage, there is an opportunity to debate and criticize the validity of the consequences and results of the model. The ultimate aim should be to achieve plausible associations between the model and the real system. At this point it should be clear that, if the conceptualization process was successful, the logical conclusions are as valid as rigorous the mathematical techniques employed, given that the model's results are a direct consequence

of the hypotheses and concepts defined in the conceptualization phase.

III. Management and Optimization

A model fulfills its objective if it is useful and fruitful for the purpose for which it was developed. The availability of such a model has then additional benefits: it allows informed management of the system and its optimization. In this vein, mathematical modeling can be combined with operations research in order to support bio-scientists in the improvement of bioprocesses with technological or biomedical purposes (Torres and Voit, 2002; Vera and Torres, 2003; Vera et al., 2010). These type of questions can be rationally answered using mathematical modeling in combination with data mining and operations research, that have been shown to be a promising approach in fields such as drug discovery (Vera et al., 2007) and operations research (Vera et al., 2010).

The optimized model, as any candidate model, should be evaluated in terms of its numerical quality in the same terms as presented above, to be a proven suitable representation of a real system (see the Parameter estimation and quality assessment section). And, as usual in these cases (see the parameter estimation and quality assessment section above) these analyses contribute to the refinement of the model through another iterative virtuous cycle (purple arrows) that superimposes to the previous one, leading to a further improved conceptual model.

CONCLUDING REMARKS

Mathematical modeling was made possible as early as the 17th century, but it is with today's techniques that it has become available to natural (and even social) scientists. There is already an ample evidence of the value and usefulness of the modeling approach in solving relevant problems in biosciences (Hübner et al., 2011; Lanza et al., 2012; Visser et al., 2014). However, in order to place modeling at the core of biological research it is necessary for the new generations of bio-scientists to be prepared to be able to build models. Currently, there are two conditions that must be met for this trend to accelerate. First, it is a matter of fact that the ecumenical nature of the training required by the modeling task in biosciences has impaired this desired evolution. The paradigm shift that involves the incorporation of the integrative approach requires shaping and expanding the training base of the new bioscience graduates with elements that include a broad and solid background in mathematics, thermodynamics, and scientific computing, among other new disciplines, in addition to the classic as chemistry, genetics and bioinformatics. Mathematical modeling of bioprocesses also has severe limitations for development and generalization because of the lack of training in math observed in many bioscience postgraduates (Watters and Watters, 2006; Koenig, 2011). It is our view that the best way to overcome this flaw is through the inclusion of two elements that are, at least not well developed in the curricula of the biosciences graduates, if not absent. One is the appropriate, and properly adapted mathematical contents, which could deal with the normally underdeveloped mathematical

thinking of the students. There is some discussion as to what contents and to what extent the biosciences students should be exposed to (Voit and Kemp, 2009). But what seems unavoidable is the fact that the biological scientist of the XXI century should be fluent not only in mathematics (in statistics and also in other mathematical concepts such as ordinary differential equations) but also in modeling techniques. Fortunately, there is an increasing awareness in this regard and some material is already available to fill this gap (Voit, 2012; Torres, 2013).

The understanding of the system through the use of the mathematical tools that allow the description and analysis of the complex systems can help to deepen the knowledge of the structure, properties and dynamic behavior of the system. However, the collaboration with experienced mathematicians is required for analyses such as the choice of the proper numerical methods, and the selection of the valid simplifications of complicated models. This is the area where most of the typical modeling projects develop: the fertile interface among established disciplines such as cellular biology, biochemistry, genetics and mathematics, and others. In this task all parties are benefited from valuable insight from the interdisciplinary experience. Modeling implies the definition of the model's objectives, and the curation of the available information. It facilitates not only the finding of previously unsuspected areas of exploration, but the proposition of new questions that were not at all evident from the reductionist approach. The systematic practice of modeling in this context also naturally facilitates the fusion of scientific disciplines; this unifying tension is felt not only among biological specialties (e.g., biochemistry, cell biology, microbiology, and genetics) but also with other (seemingly) distant ones, as operational research, computer science and mathematical analysis.

Most of the modelers are well between two extreme positions. On one side lie the idealistic ones who consider model building as a mental process in which the inductive dimension is not valued. For them the mathematical structure obtained represents reality. Opposed to this is the one with a pragmatic view, for whom the goal is to adjust the model to the actual data but without paying attention to a better understanding of reality. The right position would be that in which the model is adjusted to the data, but

reaching an understanding of the observed reality is always the aim. The optimum position of a good modeler is halfway between the most pragmatic and utilitarian view of an engineer and the more general outlook of the philosopher.

Finally, it should be noted that although the most common approach in the current biological research is the study of the design of living organisms by observing examples available in nature, there is an inductive, subsequent task that should not be forgotten. We refer to the derivation of general principles from these examples. Efforts are being carried out to gain insight into what is possible in biological design (Savageau, 1976; Alon, 2006; Salvado et al., 2011; Wolkenhauer and Green, 2013) that may lead to practical techniques for generating designs for biological systems intended to carry out particular tasks.

AUTHOR CONTRIBUTIONS

ND has contributed to the conception and design of the work and the analysis and interpretation of the relevant information and the drafting and revision of its content. Also gave the final approval of the version to be published.

GS has contributed to the conception and design of the work and the analysis and interpretation of the relevant information and the drafting and revision of its content. Also gave the final approval of the version to be published.

FUNDING

This work was funded by research grants from MINECO, Project BIO2014-54411-C2-2-R and the IMBRAIN project, REF. FP7-REGPOT-2012-CT2012-31637-IMBRAIN.

ACKNOWLEDGMENT

The authors acknowledge discussions with Dr. Daniel Guebel and Dr. Catalina Feledi for her valuable collaboration.

REFERENCES

- Allender, S., Brynle, O., Kuhlberg, J., Lowe, J., Nagorcka-Smith, P., Whelan, J., et al. (2015). Community based systems diagram of obesity causes. *PLoS ONE* 10:e0129683. doi: 10.1371/journal.pone.0129683
- Alon, U. (2006). *An Introduction to Systems Biology: Design Principles of Biological Circuits*. London: Chapman and Hall Publisher.
- Ashyraliyev, M., Fomekong-Nanfack, Y., Kaandorp, J. A., and Blom, J. G. (2008). Systems biology: parameter estimation for biochemical model. *FEBS J.* 276, 886–902. doi: 10.1111/j.1742-4658.2008.06844.x
- Banga, J. R. (2008). Optimization in computational systems biology. *BMC Syst. Biol.* 2:4. doi: 10.1186/1752-0509-2-47
- Bedau, M. A. (1999). "Can unrealistic computer models illuminate theoretical biology?," in *Proceedings of the 1999 Genetic and Evolutionary Computation Conference Workshop*, ed. A. Wu (Orlando, FL: GECCO), 20–23.
- Brauer, F., and Castillo-Chavez, C. (2010). *Mathematical Models in Population Biology and Epidemiology*. New York, NY: Springer.
- Cheong, R., Hoffmann, H., and Levchenko, A. (2008). Understanding NF- κ B signaling via mathematical modelling. *Mol. Syst. Biol.* 4, 1–11. doi: 10.1038/msb.2008.30
- ElKalaawy, N., and Wassal, A. (2015). Methodologies for the modeling and simulation of biochemical networks, illustrated for signal transduction pathways: a primer. *BioSystems* 129, 1–18. doi: 10.1016/j.biosystems.2015.01.008
- Gunawardena, J. (2011). Some lessons about models from michaelis and menten. *MBoC Perspect.* 23, 517–519. doi: 10.1091/mbc.E11-07-0643
- Hodkin, A. L., and Huxley, A. F. (1952). A quantitative description of membrane current and its application to conduction and excitation in nerve. *J. Physiol.* 117, 500–544. doi: 10.1113/jphysiol.1952.sp004764
- Hsiung, W., Feng, W., Wang, S., and Chang, M. S. (2008). Dynamic sensitivity analysis of biological systems. *BMC Bioinformatics* 9:S17. doi: 10.1186/1471-2105-9-S12-S17
- Hübner, K., Sahle, S., and Kummer, U. (2011). Applications and trends in systems biology in biochemistry. *FEBS J.* 278, 2767–2857. doi: 10.1111/j.1742-4658.2011.08217.x
- Janes, K. A., and Yaffe, M. B. (2006). Data-driven modelling of signal-transduction networks. *Nat. Rev. Mol. Cell Biol.* 7, 820–828. doi: 10.1038/nrm2041
- Koenig, J. (2011). A Survey of the Mathematics Landscape within Bioscience Undergraduate and Postgraduate UK Higher Education. UK Centre for Bioscience, Higher Education Academy.

- Krishna, S., Jensen, M. H., and Sneppen, K. (2006). Minimal model of spiky oscillations in NF- κ B signalling. *Proc. Natl. Acad. Sci. U.S.A.* 103, 10840–10845. doi: 10.1073/pnas.0604085103
- Kuhn, T. S. (1962). *The Structure of Scientific Revolutions*. Chicago: University of Chicago Press.
- Kuhn, T. S. (1965). Logic of discovery or psychology of research. *Crit. Growth Knowl.* 1970, 1–23.
- Lakatos, I. (1970). Falsification and the methodology of scientific research programs. *Crit. Growth Knowl.* 1970, 91–95.
- Lanza, A. M., Crook, N. C., and Alper, H. S. (2012). Innovation at the intersection of synthetic and systems biology. *Curr. Opin. Biotechnol.* 23, 712–717. doi: 10.1016/j.copbio.2011.12.026
- Lotka, A. L. (1920). Analytical note on certain rhythmic relations in organic systems. *Proc. Natl. Acad. Sci. U.S.A.* 6, 410–415. doi: 10.1073/pnas.6.7.410
- MacDonald, G., Cuellar, C. B., and Foll, C. V. (1968). The dynamics of malaria. *Bull. World Health Organ.* 38, 743–755.
- Marrink, S. J., and Tieleman, D. P. (2013). Perspective on the Martini model. *Chem. Soc. Rev.* 2013:6801. doi: 10.1039/c3cs60093a
- Medio, A. (2006). *Mathematical Models in Economy: Mathematical Models*, Vol. 3. Paris: Encyclopedia of Life Support Systems, 1–3.
- Michaelis, L., and Menten, M. L. (1913). Die Kinetik der Invertinwirkung. *Biochem. Z.* 49, 333–369.
- Minegishi, S., and Thiel, D. (2000). System dynamics modelling and simulation of a particular food supply chain. *Simul. Pract. Theory* 8, 321–339. doi: 10.1016/S0928-4869(00)00026-4
- Monod, J., Wyman, J., and Changeux, J. (1965). On the nature of allosteric transitions: a plausible model. *J. Mol. Biol.* 12, 88–118. doi: 10.1016/S0022-2836(65)80285-6
- Ni, T. C., and Savageau, M. A. (1996a). Application of biochemical systems theory to metabolism in human red blood cells. Signal propagation and accuracy of representation. *J. Biol. Chem.* 271, 7927–7941. doi: 10.1074/jbc.271.14.7927
- Ni, T. C., and Savageau, M. A. (1996b). Model assessment and refinement using strategies from biochemical systems theory: application to metabolism in human red blood cells. *J. Theor. Biol.* 179, 329–368. doi: 10.1006/jtbi.1996.0072
- Okamoto, M., and Savageau, M. (1984). Integrated function of a kinetic proofreading mechanism: steady-state analysis testing internal consistency of data obtained in vivo and in vitro and predicting parameter values. *Biochemistry* 23, 1701–1709. doi: 10.1021/bi00303a019
- Okamoto, M., and Savageau, M. (1986). Integrated function of a kinetic proofreading mechanisms: double-stage proofreading by isoleucyl-tRNA synthetase. *Biochemistry* 25, 1969–1975. doi: 10.1021/bi00356a020
- Park, L. J., Park, C. H., Park, C., and Lee, T. (1997). Application of genetic algorithms to parameter estimation of bioprocesses. *Med. Biol. Eng. Comput.* 35, 47–49. doi: 10.1007/BF02510391
- Polisety, P. K., and Voit, E. O. (2006). Identification of metabolic system parameters using global optimization methods. *Theor. Biol. Med. Modell.* 3:4. doi: 10.1186/1742-4682-3-4
- Popper, K. (1935). *The Logic of Scientific Discovery*. Vienna: Verlag von Julius Springer.
- Prigogine, I. (1961). *Introduction to Thermodynamics of Irreversible Processes*. New York, NY: Interscience Publisher.
- Ross, R. (1915). Some a priori pathometric equations. *Br. Med. J.* 1, 546–547. doi: 10.1136/bmj.1.2830.546
- Salvado, B., Karathia, H., Chimenos, A. U., Vilaprinyo, E., Omholt, S., Sorribas, A., et al. (2011). Methods for and results from the study of design principles in molecular systems. *Math. Biosci.* 231, 3–18. doi: 10.1016/j.mbs.2011.02.005
- Santos, G., and Torres, N. V. (2013). New Targets for Drug Discovery against Malaria. *PLoS ONE* 8:e59968. doi: 10.1371/journal.pone.0059968
- Savageau, M. A. (1971). Parameter sensitivity as a criterion for evaluating and comparing the performance of biochemical systems. *Nature* 229, 542–544. doi: 10.1038/229542a0
- Savageau, M. A. (1976). *Biochemical Systems Analysis: A Study of Function and Design in Molecular Biology*. Boston, MA: Addison-Wesley.
- Schank, J. C. (2008). The development of locomotor kinematics in neonatal rats: an agent-based modelling analysis in group and individual contexts. *J. Theor. Biol.* 254, 826–842. doi: 10.1016/j.jtbi.2008.07.024
- Segre, D., Vitkup, D., and Church, G. M. (2002). Analysis of optimality in natural and perturbed metabolic networks. *Proc. Natl. Acad. Sci. U.S.A.* 99, 15112–15117. doi: 10.1073/pnas.232349399
- Torres, N. V. (2013). Introducing systems biology to bioscience students through mathematical modelling: a practical module. *Biosci. Educ.* 21, 54–63. doi: 10.11120/beej.2013.00012
- Torres, N. V., and Voit, E. O. (2002). *Pathway Analysis and Optimization in Metabolic Engineering*. Cambridge: Cambridge University Press.
- van Riel, N. A. W. (2006). Dynamic modelling and analysis of biochemical networks: mechanism-based models and model-based experiments. *Brief. Bioinform.* 7, 364–374. doi: 10.1093/bib/bbl040
- Vera, J., Bachmann, J., Pfeifer, A. C., Becker, V., Hormiga, J. A., and Torres Darias, N. V. (2008). A systems biology approach to analyse amplification in the JAK2-STAT5 signalling pathway. *BMC Syst. Biol.* 2:38. doi: 10.1186/1752-0509-2-38
- Vera, J., Curto, R., Cascante, M., and Torres, N. V. (2007). Detection of potential enzyme targets by metabolic modelling and optimization. *Appl. Simple Enzymopathy Bioinform.* 23, 2281–2289. doi: 10.1093/bioinformatics/btm326
- Vera, J., González-Alcón, C., Marín-Sanguino, A., and Torres, N. V. (2010). Optimization of biochemical systems through mathematical programming: methods and applications. *Comput. Operat. Res.* 37, 1427–1438. doi: 10.1016/j.cor.2009.02.021
- Vera, J., and Torres, N. V. (2003). MetMAP: an integrated Matlab™ package for analysis and optimization of metabolic systems. *In Silico Biol.* 4:10. doi: 10.1016/j.mechrescom.2011.03.011
- Visser, S. A. G., Alwis, D. P., Kerbusch, T., Stone, J. A., and Allerheiligen, S. R. B. (2014). Implementation of quantitative and systems pharmacology in large pharma. *CPT Pharm. Syst. Pharmacol.* 3:e142. doi: 10.1038/psp.2014.40
- Voit, E. (1992). Optimization in integrated biochemical systems. *Biotechnol. Bioeng.* 40, 572–582. doi: 10.1002/bit.260400504
- Voit, E. O. (2000). *Computational Analysis of Biochemical Systems. A Practical Guide for Biochemists, and Molecular Biologist*. Cambridge: Cambridge University Press.
- Voit, E. O. (2012). *A First Course in Systems Biology*. New York, NY: Garland Science.
- Voit, E. O., and Kemp, M. L. (2009). So, you want to be a systems biologist? Determinants for creating graduate curricula in systems biology. *IET Syst. Biol.* 5, 70–79. doi: 10.1049/iet-syb.2009.0071
- Volterra, V. (1926). Variazioni e Fluttuazioni del Numero d'Individui in Specie Animali Conventi. *Mem. Acad. Lincei Roma.* 2, 31–113.
- Watters, D. J., and Watters, J. J. (2006). Student understanding of pH. “I Don’t know what the log actually is, I only know where the button is on my calculator”. *Biochem. Mol. Biol. Educ.* 34, 278–284. doi: 10.1002/bmb.2006.494034042628
- Weidlich, W. (2003). Sociodynamics—a systematic approach to mathematical modelling in the social sciences. *Chaos Solitons Fract.* 18, 431–437. doi: 10.1142/S0219477503001294
- Wimsatt, W. C. (1987). “Re-Engineering philosophy for limited beings,” in *Neutral Models in Biology*, eds M. H. Nitecki and A. Hoffman (New York, NY: Oxford University Press), 23–55.
- Wolkenhauer, O., and Green, S. (2013). The search for organizing principles as a cure against reductionism in systems medicine. *FEBS J.* 280, 5938–5948. doi: 10.1111/febs.12311
- Wu, W. H., Wang, F. S., and Chang, M. S. (2008). Dynamic sensitivity analysis of biological systems. *BMC Bioinform.* 9:S17. doi: 10.1186/1471-2105-9-S12-S17

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

The reviewer Ester Vilaprinyo and handling Editor Rui Alves declare that despite their previous collaborations the review process was conducted objectively.

Copyright © 2015 Torres and Santos. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



Design Space Toolbox V2: Automated Software Enabling a Novel Phenotype-Centric Modeling Strategy for Natural and Synthetic Biological Systems

Jason G. Lomnitz¹ and Michael A. Savageau^{1,2*}

¹ Department of Biomedical Engineering, University of California, Davis, Davis, CA, USA, ² Department of Microbiology and Molecular Genetics, University of California, Davis, Davis, CA, USA

OPEN ACCESS

Edited by:

Julio Vera González,
University Hospital Erlangen, Germany

Reviewed by:

Osbaldo Resendis-Antonio,
Instituto Nacional de Medicina
Genómica, Mexico
Alessandro Giuliani,
Istituto Superiore di Sanità, Italy

*Correspondence:

Michael A. Savageau
masavageau@ucdavis.edu

Specialty section:

This article was submitted to
Systems Biology,
a section of the journal
Frontiers in Genetics

Received: 01 March 2016

Accepted: 07 June 2016

Published: 12 July 2016

Citation:

Lomnitz JG and Savageau MA (2016)
Design Space Toolbox V2: Automated
Software Enabling a Novel
Phenotype-Centric Modeling Strategy
for Natural and Synthetic Biological
Systems. *Front. Genet.* 7:118.
doi: 10.3389/fgene.2016.00118

Mathematical models of biochemical systems provide a means to elucidate the link between the genotype, environment, and phenotype. A subclass of mathematical models, known as mechanistic models, quantitatively describe the complex non-linear mechanisms that capture the intricate interactions between biochemical components. However, the study of mechanistic models is challenging because most are analytically intractable and involve large numbers of system parameters. Conventional methods to analyze them rely on local analyses about a nominal parameter set and they do not reveal the vast majority of potential phenotypes possible for a given system design. We have recently developed a new modeling approach that does not require estimated values for the parameters initially and inverts the typical steps of the conventional modeling strategy. Instead, this approach relies on architectural features of the model to identify the phenotypic repertoire and then predict values for the parameters that yield specific instances of the system that realize desired phenotypic characteristics. Here, we present a collection of software tools, the Design Space Toolbox V2 based on the System Design Space method, that automates (1) enumeration of the repertoire of model phenotypes, (2) prediction of values for the parameters for any model phenotype, and (3) analysis of model phenotypes through analytical and numerical methods. The result is an enabling technology that facilitates this radically new, phenotype-centric, modeling approach. We illustrate the power of these new tools by applying them to a synthetic gene circuit that can exhibit multi-stability. We then predict values for the system parameters such that the design exhibits 2, 3, and 4 stable steady states. In one example, inspection of the basins of attraction reveals that the circuit can count between three stable states by transient stimulation through one of two input channels: a positive channel that increases the count, and a negative channel that decreases the count. This example shows the power of these new automated methods to rapidly identify behaviors of interest and efficiently predict parameter values for their realization. These tools may be applied to understand complex natural circuitry and to aid in the rational design of synthetic circuits.

Keywords: biochemical systems theory, gene regulatory circuits, System Design Space, synthetic biology, code:python

INTRODUCTION

One of the current challenges in biology is to understand the mapping between a particular genotype and a particular phenotype in the context of a specific environment. The overarching goal, with important consequences to science, biotechnology, and medicine, is to predict the phenotypes that arise from changes in the genotype, the environment or both. However, phenotypes emerge from complex systems of biochemical components that are inherently non-linear, with intricate interactions that cannot be understood through intuition. Thus, it is no surprise that the “genotype-to-phenotype” problem is commonly regarded as one of the grand challenges in modern biology (Brenner, 2000).

A prominent approach to address this difficult challenge is to formulate mathematical models and analyze them to gain detailed insight into the design principles underlying biochemical mechanisms (Savageau, 2009). However, most mathematical models of biological mechanisms are analytically intractable and therefore their study tends to be uniquely tailored and limited in scope.

The conventional strategy for the analysis of mathematical models of non-linear phenomena typically begins with a combination of experimentally measured values for a subset of system parameters and mathematically estimated values for the (often many) remaining parameters (e.g., Sun et al., 2012). The result is an established set of parameter values that serves as the focus for analyses that provide local information regarding system behavior. Thus, this conventional strategy might be termed a parameter-centric approach. When the number of parameters is small and the data is rich, this approach can be very successful. However, it is most often the case that the available experimental data is limited and the number of system parameters is large. Consequently, this approach has severe limitations when attempting to discover the repertoire of potential phenotypes latent in a particular system design.

We have recently developed a radically new modeling strategy that—unlike the parameter-centric approach—does not depend on specific values for the parameters (Lomnitz and Savageau, 2015). This new, phenotype-centric, approach builds on and extends the System Design Space method (Savageau et al., 2009; Fasani and Savageau, 2010; Lomnitz and Savageau, 2013) by (1) enumerating the repertoire of model phenotypes latent in a particular system design, (2) identifying phenotypes that exhibit characteristics of interest, and (3) predicting parameter values for the realization of a specific instance of the system exhibiting the characteristics of interest (Lomnitz and Savageau, 2015).

Here, we present a collection of software tools, the Design Space Toolbox V2, that automates the most difficult steps of this strategy. These software tools build on a previous iteration, the Design Space Toolbox for Matlab®, that formalized automatic construction of the design space for biochemical systems (Fasani and Savageau, 2010). The new tools we present here automate the deconstruction of a model into qualitatively distinct phenotypes—thereby automatically enumerating the phenotypic repertoire of the system (Lomnitz and Savageau, 2015). These

tools improve upon the previous iteration by addressing key bottlenecks and expanding upon its capabilities through new technologies that enable analyses not previously possible.

The most important contributions from these new tools include (1) a complete redesign for improved resource management and parallelization of the algorithms for concurrent analysis of model phenotypes; (2) automation of the analysis of local stability through an expansion of the analytical capabilities of the tools; (3) automation of the prediction of parameter values for phenotypes of the system, and (4) automating the co-localization of cases to determine the simultaneous realization and visualization of ensembles of model phenotypes (Lomnitz and Savageau, 2015).

We illustrate the capabilities of these new tools and the thought process guiding the new modeling approach by means of an example. Although we use a model with specified mechanisms for illustrative purposes, in practice one will undoubtedly have only partial information about the underlying mechanisms and one must fill in the missing information by making hypotheses that need to be tested. In our method one need only postulate the architectural information: the topology, signs, and stoichiometry of the interactions. As we have discussed elsewhere (Lomnitz and Savageau, 2015), these are the features of a model that are most readily obtained by experiment or by means of sampling a small number of integers. The more difficult values to determine are rate constants and binding constants, which our method handles automatically in the process of testing the hypotheses. Our method allows for the efficient testing of alternative models by automatic enumeration of the phenotypic repertoire and prediction of model parameters without numerical estimation or sampling of a high-dimensional parameter space. In a recent application we tested 40 different models (hypotheses) and found only five that were consistent with the experimental data (Lomnitz and Savageau, in review).

Following the detailed illustration of the methods, we apply them to a mechanistic model for a new synthetic gene circuit, proposed here, that can exhibit multi-stability involving up to four steady states. Furthermore, we show that this circuit can alternate between three distinct states in a step-wise fashion through the transient stimulation in one of two input channels—a positive channel that results in forward transitions through the three states and a negative channel that results in reverse transitions through the three states. In this way, we describe a genetic counter that can count between three states that—unlike other genetic counters that can count transiently (e.g., see Friedland et al., 2009)—can retain its count indefinitely.

This example shows the power of these new automated tools to provide insight into the underlying design principles of a system involving complex non-linear interactions that are ubiquitous in biology. We also have shown that these tools are useful for designing novel synthetic gene circuits that may be important for a variety of applications from biotechnology (e.g., Martin et al., 2003) to medicine (e.g., Ro et al., 2006), and for gaining insights into more complex natural circuitry (e.g., Benner and Sismour, 2005; Stricker et al., 2008; Mukherji and

van Oudenaarden, 2009; Tigges et al., 2009; Kim and Forger, 2012).

BACKGROUND

In this section, we review key concepts and definitions from the System Design Space methodology (Savageau et al., 2009) that deconstructs systems based on differences in phenotypes (Lomnitz and Savageau, 2014). As a vehicle to facilitate presentation of the basic concepts, we apply the System Design Space method to a simple example involving a single gene regulator that is autogenously controlled via a positive feedback loop that exhibits the potential for bistability. In a later section, we build on this simple example to show how our automated tools can be applied to a more complex circuit.

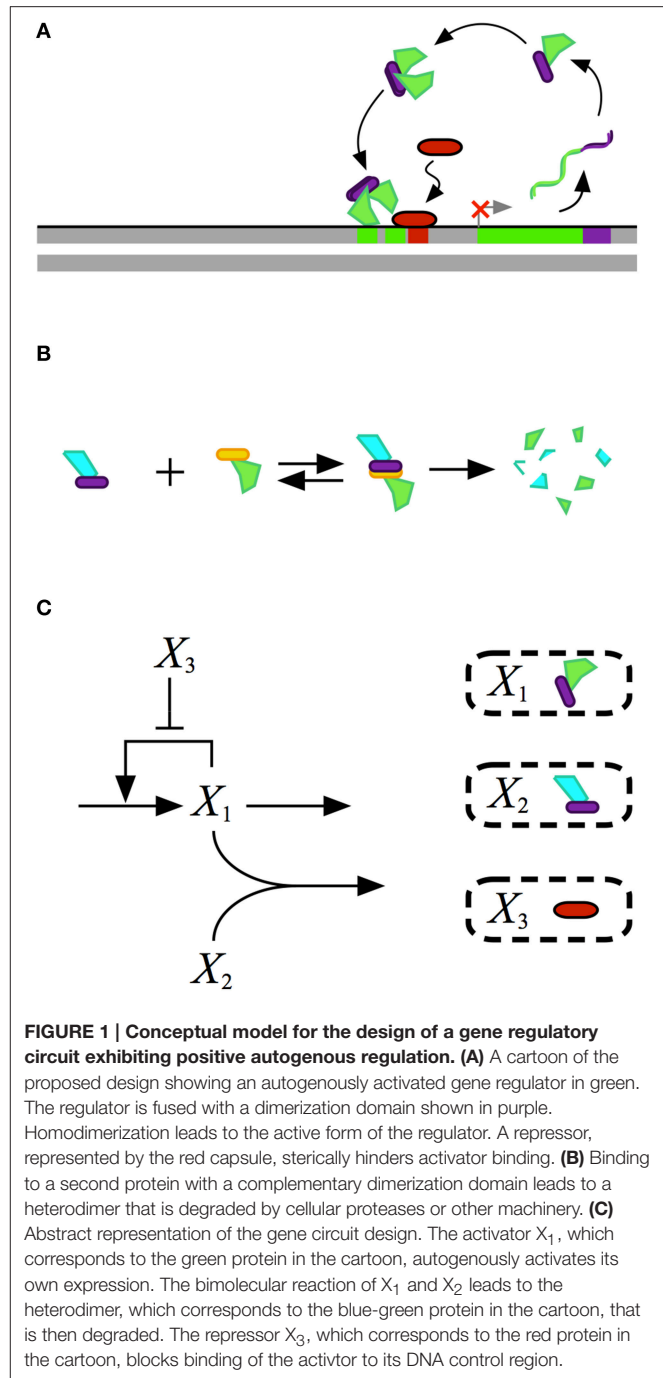
We analyze the system by (1) formulating a mechanistic model of a simple biochemical system; (2) recasting the model into the generic Generalized Mass Action (GMA) form; (3) constructing the design space for the recast GMA-System; (4) enumerating the phenotypic repertoire of the model; and (5) analyzing model phenotypes to identify their phenotypic characteristics.

Formulating a Mechanistic Model of a Biochemical System

Mathematical modeling of biochemical phenomena usually begins with the synthesis of available knowledge from the literature and experimental data that together provide a foundation for generating a particular hypothesis. The hypothesis is usually represented by a conceptual model that contains qualitative information regarding the key components and their interactions, typically visualized using some sort of diagram. An example of a conceptual model for a simple gene regulatory circuit is represented in **Figure 1**.

Once the qualitative aspects of a system and its interactions have been realized in a conceptual model, we formulate mathematical models by hypothesizing specific biochemical mechanisms involving the elementary rate laws of chemical kinetics and the rational function rate laws of biochemical kinetics (Lomnitz and Savageau, 2013). The result is a system of non-linear differential equations that is analytically intractable in all but the simplest cases (Lomnitz and Savageau, 2014).

In general, the exponents in the power laws that characterize classical chemical kinetics are small integer values, as are the exponents in the rational functions that characterize classical biochemical kinetics. In the case of an enzymatic reaction, the largest exponent in the rate law is equal to the number of reactant binding sites on the enzyme (Wyman, 1964), and this is typically equal to the number of subunits in a multimeric protein (Monod et al., 1965). In the case of a regulator that is a multimeric DNA binding protein, the largest exponent is equal to the number of subunits in the regulator molecule multiplied by the number of specific sites on the DNA to which it binds. Experimental evidence indicates that regulators function as multimeric, typically dimeric, molecules that bind a single recognition site, or possibly a small number of such sites cooperatively, for each transcriptional unit controlled (Mandal



et al., 1990; Kim and Little, 1992). If the mechanisms in the model are known, then the exponents will be known; if one has to hypothesize a mechanism, then one has only to sample a small number of fixed integer values for the exponents to characterize the model.

The aspects of a mathematical model that remain fixed for a particular mechanism—independent of the specific values for the parameters that characterize a particular instantiation—are defined as its *architectural features* (Lomnitz and Savageau, 2015). These features include (a) the network topology of interactions,

(b) the signs of the interactions, and (c) the number of binding sites involved in the interactions that in turn manifests itself in the exponents found in the power laws of chemical kinetics and in the rational functions of biochemical kinetics, which, as noted above, are fixed integers for a particular mechanism.

A mathematical model for the conceptual system shown in **Figure 1** is represented by the following ordinary differential equation (ODE),

$$\frac{dX_1}{dt} = \alpha_1 \left[\frac{1 + \rho_1 \left(\frac{X_1}{K_1} \right)^2 + \frac{X_3}{K_3}}{1 + \left(\frac{X_1}{K_1} \right)^2 + \frac{X_3}{K_3}} \right] - \beta_1 X_1 - k X_1 X_2 \quad (1)$$

where X_1 represents the dependent activator protein; X_2 represents a protein with a complimentary heterodimerization domain, X_3 represents the repressor protein, and their values are treated as independent variables; α_1 represents the basal level of expression for the synthesis of X_1 ; β_1 represents the first-order rate constant for the loss of X_1 by dilution due to exponential growth; ρ_1 represents the capacity for activation of X_1 synthesis; K_1 represents the concentration of X_1 for the half-maximal rate of synthesis; K_3 represents the concentration of X_3 that results in half-maximal repression; and k represents the rate constant for X_1 - X_2 heterodimer formation.

This model makes conventional assumptions found in the literature regarding the mechanisms for the control of transcription, and for the translation and loss of stable proteins by dilution due to exponential growth. However, if the mechanisms were unknown, one could postulate alternative mechanisms, as outlined in the Introduction, and test the hypothesized models against experimental data.

Recasting Equations into a Generic Form

The System Design Space method provides a novel approach to deconstruct mathematical models of biochemical systems (Savageau et al., 2009). At its core, this approach utilizes an innovative definition for model phenotypes that is based on dominant processes that produce sub-systems exhibiting qualitatively-distinct behavior (Savageau et al., 2009).

In order to apply the System Design Space method, the system must first be recast into the canonical GMA form involving a system of differential equations plus algebraic constraints expressed mathematically as,

$$\frac{dX_1}{dt} = \sum_{k=1}^{P_1} \alpha_{1k} \prod_{j=1}^{n+m} X_j^{g_{1jk}} - \sum_{k=1}^{Q_1} \beta_{1k} \prod_{j=1}^{n+m} X_j^{h_{1jk}} \quad (2)$$

$$\vdots$$

$$\frac{dX_{n_t}}{dt} = \sum_{k=1}^{P_{n_t}} \alpha_{n_t k} \prod_{j=1}^{n+m} X_j^{g_{n_t j k}} - \sum_{k=1}^{Q_{n_t}} \beta_{n_t k} \prod_{j=1}^{n+m} X_j^{h_{n_t j k}}$$

$$0 = \sum_{k=1}^{P_{n_t+1}} \alpha_{(n_t+1)k} \prod_{j=1}^{n+m} X_j^{g_{(n_t+1)jk}} - \sum_{k=1}^{Q_{n_t+1}} \beta_{(n_t+1)k} \prod_{j=1}^{n+m} X_j^{h_{(n_t+1)jk}}$$

$$\vdots$$

$$0 = \sum_{k=1}^{P_n} \alpha_{nk} \prod_{j=1}^{n+m} X_j^{g_{nj k}} - \sum_{k=1}^{Q_n} \beta_{nk} \prod_{j=1}^{n+m} X_j^{h_{nj k}} \quad (3)$$

where n_t represents the number of dynamic variables; n_c represents the number of auxiliary variables; $n = n_t + n_c$ represents the number of dependent variables; m represents the number of independent variables; α_{ik} represents the rate constant for the k -th positive term of the i -th equation; β_{ik} represents the rate constant for the k -th negative term of the i -th equation; P_i and Q_i represent the number of positive and negative terms for the i -th equation, respectively; g_{ijk} and h_{ijk} represents the kinetic order for the influence of the j -th variable on the k -th positive and negative term of the i -th equation, respectively; X_j represent the j -th variable such that the first n_t variables are the dynamic variables, the second n_c are the auxiliary variables and the last m variables are the independent variables.

Mechanistic models of biochemical phenomena can be recast exactly into this form by following a well-defined series of steps (Savageau and Voit, 1987). Furthermore, for most biochemical systems the recasting process is straight-forward and involves five simple steps: (1) expanding terms in the numerator by multiplying through by common factors; (2) defining auxiliary variables for each denominator that has multiple terms; (3) rearranging terms in the equation for the auxiliary variables so that the left-hand side is equal to 0; (4) substituting the auxiliary variables for the corresponding denominators; and (5) defining a new system of differential-algebraic equations involving the modified differential equations and the algebraic equations for the auxiliary variables.

We illustrate the process by recasting into the GMA form Equation (1), which involves a typical rational function from biochemical kinetics.

Step 1. Expand the numerator of the equation for X_1 by multiplying through by the α parameter.

$$\frac{dX_1}{dt} = \frac{\alpha_1 + \alpha_1 \rho_1 \left(\frac{X_1}{K_1} \right)^2 + \alpha_1 \frac{X_3}{K_3}}{1 + \left(\frac{X_1}{K_1} \right)^2 + \frac{X_3}{K_3}} - \beta_1 X_1 - k X_1 X_2 \quad (4)$$

Step 2. Define an auxiliary variable, X_{100} , equal to the expression in the denominator.

$$X_{100} = 1 + \left(\frac{X_1}{K_1} \right)^2 + \frac{X_3}{K_3} \quad (5)$$

Step 3. Rearrange terms in the new equation so that the left-hand side of the equation is equal to 0.

$$0 = 1 + \left(\frac{X_1}{K_1} \right)^2 + \frac{X_3}{K_3} - X_{100} \quad (6)$$

Step 4. Substitute the auxiliary variable for the denominator of the equation from Step 1.

$$\frac{dX_1}{dt} = \alpha_1 X_{100}^{-1} + \alpha_1 \rho_1 X_1^2 K_1^{-2} X_{100}^{-1} + \alpha_1 X_3 K_3^{-1} X_{100}^{-1} - \beta_1 X_1 - k X_1 X_2 \quad (7)$$

Step 5. Define a new system by including the algebraic constraint from Step 3.

$$\frac{dX_1}{dt} = \alpha_1 X_{100}^{-1} + \alpha_1 \rho_1 X_1^2 K_1^{-2} X_{100}^{-1} + \alpha_1 X_3 K_3^{-1} X_{100}^{-1} - \beta_1 X_1 - k X_1 X_2 \quad (8)$$

$$0 = 1 + X_1^2 K_1^{-2} + X_3 K_3^{-1} - X_{100} \quad (9)$$

The result is a differential-algebraic system in a generic form consisting of linear combinations of non-linear terms having a very specific structure (products of power laws) that is capable of representing a broad range of non-linear systems (Lomnitz and Savageau, 2013). It should be noted that from a mathematical perspective both independent variables and system parameters are treated equally within the context of the System Design Space method (Fasani and Savageau, 2010); thus, in this article we use the terms independent variables and parameters interchangeably to refer to their combined set.

Mathematical Definition of Qualitatively-Distinct Phenotypes

The system of equations in the GMA form can be analyzed by using the novel System Design Space method. This method deconstructs complex non-linear systems into a finite number of qualitatively-distinct, non-linear, sub-systems (S-Systems). The qualitatively-distinct phenotypes are mathematically defined in terms of these sub-system equations (Savageau et al., 2009; Lomnitz and Savageau, 2013) and their system behavior is tractable for a variety of system properties (Savageau, 2009; Voit, 2013).

Grouping of Terms

The mathematical definition of qualitatively-distinct phenotype originates from the structure of the GMA-system. Inspection of this generalized form, shown in Equations (2) and (3), reveals a regular structure: for any i -th equation, the right-hand side is a sum of P_i positive terms and Q_i negative terms. Therefore, a system will have a *system signature* that involves a listing of the number of positive and negative terms, i.e., $(P_1 Q_1 P_2 Q_2 \dots P_n Q_n)$ (Savageau et al., 2009; Fasani and Savageau, 2010; Lomnitz and Savageau, 2015).

Dominant Terms

At any given point in the combined variable and parameter space of the system, where each variable and parameter has a specific value, the magnitude of the terms in each equation can be quantified and the terms with a given sign can be ranked based on their relative magnitude. A *dominant term* is defined as the largest term of a given sign for an equation of the GMA-system; and the dominant terms with positive and negative signs are the *dominant positive term* and the *dominant negative term*, respectively (Savageau et al., 2009).

The dominant terms can be uniquely identified based on the index in their corresponding summations. The combination of indices for dominant terms for all the equations yields a unique *case signature* that involves a listing of indices of dominant positive and dominant negative terms in order, i.e.,

$[p_1 q_1 p_2 q_2 \dots p_n q_n]$ (Savageau et al., 2009; Fasani and Savageau, 2010; Lomnitz and Savageau, 2015), where p_i and q_i are the indices of the dominant positive term and dominant negative term of the i -th equation, respectively. Note that the system signature (surrounded by parentheses) is differentiated from the case signatures (surrounded by square brackets).

Dominant S-Systems

Any point in the variable plus parameter space has a corresponding combination of dominant terms. Because the possible combinations of dominant terms are finite, with the maximum determined by $\prod_i^n P_i Q_i$, this partitions the space into a set of discrete “chunks” that are identifiable by their unique case signature (Savageau et al., 2009; Fasani and Savageau, 2010; Lomnitz and Savageau, 2013). Each discrete chunk has a unique combination of dominant terms and, by retaining only the dominant terms and neglecting the non-dominant terms, we can define a *dominant sub-system* that is characteristic of a particular “chunk.”

The dominant sub-systems, defined by retaining only the dominant terms, have a very special structure. These equations are *S-Systems* that have a single positive term and a single negative term that are products of power laws given by the following equations,

$$\frac{dX_1}{dt} = \alpha_{1p_1} \prod_{j=1}^{n+m} X_j^{g_{1jp_1}} - \beta_{1q_1} \prod_{j=1}^{n+m} X_j^{h_{1jq_1}} \quad (10)$$

$$\vdots$$

$$\frac{dX_{n_t}}{dt} = \alpha_{1p_{n_t}} \prod_{j=1}^{n+m} X_j^{g_{1jp_{n_t}}} - \beta_{1q_{n_t}} \prod_{j=1}^{n+m} X_j^{h_{1jq_{n_t}}}$$

$$0 = \alpha_{1p_{(n_t+1)}} \prod_{j=1}^{n+m} X_j^{g_{1jp_{(n_t+1)}}} - \beta_{1q_{(n_t+1)}} \prod_{j=1}^{n+m} X_j^{h_{1jq_{(n_t+1)}}}$$

$$\vdots$$

$$0 = \alpha_{1p_n} \prod_{j=1}^{n+m} X_j^{g_{1jp_n}} - \beta_{1q_n} \prod_{j=1}^{n+m} X_j^{h_{1jq_n}} \quad (11)$$

The steady-state equations for S-Systems are non-linear but tractable because they become linear when transformed into logarithmic coordinates (Savageau, 2009; Voit, 2013).

Dominance Conditions

If we had to sample the full $(n + m)$ -dimensional space of a system—where n is the number of dependent variables plus auxiliary variables and m is the number of independent variables plus parameters—to identify the regions associated with each qualitatively-distinct phenotype, the usefulness of this approach would be limited. However, the fact that each term is a product of power laws makes possible more extensive analysis of the conditions that partition the continuous variable and parameter space into discrete regions that define the design space of a system.

Dominance can be expressed mathematically through a series of inequalities. The inequalities for the dominant terms of the i -th

equation are given by,

$$\alpha_{ip_i} \prod_{j=1}^{n+m} X_j^{g_{ijp_i}} > \alpha_{ik} \prod_{j=1}^{n+m} X_j^{g_{ijk}} \forall k = \{1, 2, 3, \dots, P_i | k \neq p_i\} \quad (12)$$

$$\beta_{iq_i} \prod_{j=1}^{n+m} X_j^{h_{ijq_i}} > \beta_{ik} \prod_{j=1}^{n+m} X_j^{h_{ijk}} \forall k = \{1, 2, 3, \dots, P_i | k \neq q_i\} \quad (13)$$

which can be transformed to yield a series of linear inequalities in the logarithm of the variables,

$$\log \alpha_{ip_i} + \sum_{j=1}^{n+m} g_{ijp_i} \log X_j > \log \alpha_{ik} + \sum_{j=1}^{n+m} g_{ijk} \log X_j \quad \forall k = \{1, 2, 3, \dots, P_i | k \neq p_i\} \quad (14)$$

$$\log \beta_{iq_i} + \sum_{j=1}^{n+m} h_{ijq_i} \log X_j > \log \beta_{ik} + \sum_{j=1}^{n+m} h_{ijk} \log X_j \quad \forall k = \{1, 2, 3, \dots, P_i | k \neq q_i\} \quad (15)$$

Because these inequalities are linear, they have the following characteristics: (1) each condition defines a half-space of the $(n + m)$ -dimensional space (i.e., half the $(n + m)$ -dimensional space); and (2) the intersection of all the half-spaces yields either (a) an $(n + m)$ -dimensional *dominance polytope* (i.e., there is a feasible region for the phenotype in the state plus parameter space) or (b) a null region (i.e., there is no feasible region for the phenotype anywhere in the combined state plus parameter space). The validity of the dominance polytope can be determined very efficiently and is typically the first phase of a linear programming problem (Vanderbei, 1996).

Boundary Conditions

The steady-state solution of a dominant S-System is linear in logarithmic coordinates (Savageau, 2009). The boundary conditions for validity of the corresponding phenotype are obtained by substituting the linear solution for the steady state into the linear dominance conditions, to yield boundaries for the dominant sub-system that are linear in logarithmic space (Savageau et al., 2009; Fasani and Savageau, 2010). Each boundary condition defines an m -dimensional half-space and the intersection of these half-spaces yields an m -dimensional phenotypic polytope.

From a geometric perspective, the steady-state solution defines an m -dimensional *solution hyperplane* that “cuts” through the $(n + m)$ -dimensional *dominance polytope*. The boundary conditions are the edges at the intersection between the solution hyperplane and the dominance polytope, which yields the phenotypic polytope of the system.

However, the boundary conditions may not necessarily yield a feasible region because of two reasons: (a) the dominant S-System is underdetermined and has no steady-state solution or (b) the solution hyperplane and the dominance polytope do not intersect anywhere in the $(n + m)$ -dimensional space. The validity of

the feasible region can be determined in the same way as the validity of the dominance polytope by using linear programming methods (Fasani and Savageau, 2010).

Qualitatively Distinct Phenotypes

The dominant S-Systems capture the behavior of the system’s dominant processes contributing to the synthesis and loss for each species. These non-linear sub-systems, with particular phenotypic characteristics, capture the dominant behaviors of the full system. These sub-systems are valid representations of the system behavior within mathematically defined boundaries that are analytically determined by the system equations themselves. The combination of a characteristic sub-system and mathematically defined boundaries partitions parameter space into a finite number of regions where the system behavior has a series of characteristic traits. The result is a mathematical definition for qualitatively-distinct phenotypes that is based on the processes of a given system that are dominant in a particular context (Savageau et al., 2009; Lomnitz and Savageau, 2015).

Phenotypic Repertoire

The phenotypic repertoire is defined as the collection of qualitatively-distinct phenotypes (valid phenotypic polytopes), integrated into a space-filling structure known as the system design space (Lomnitz and Savageau, 2015).

DESIGN SPACE TOOLBOX V2

It is widely recognized that the phenotype-to-genotype challenge is difficult in large part because the tools available for the analysis of non-linear systems have little power to explore the global landscape of system behavior. Thus, most analyses rely on estimating values for the parameters and analyzing the system at a local level. The System Design Space method addresses some of these limitations by providing detailed information about the system behavior from a global perspective (Lomnitz and Savageau, 2014). It does this by enumerating the repertoire of a system’s qualitatively-distinct phenotypes and identifying a subset of phenotypes of interest. It achieves this by deconstructing intractable non-linear systems into tractable non-linear sub-systems that can be reassembled to define a system’s design space (Savageau et al., 2009).

We have recently applied this methodology to a variety of biochemical systems that exhibit rich behaviors including multi-stability (Savageau and Fasani, 2009; Martínez-Antonio et al., 2012; Fasani and Savageau, 2013) and oscillations (Lomnitz and Savageau, 2013, 2014, 2015). Other examples involve natural gene circuits that play crucial roles in the transitions between alternative modes of biological operation [e.g., aerobic to anaerobic growth (Tolla and Savageau, 2010, 2011; Tolla et al., 2015), growth phase transitions (Martínez-Antonio et al., 2012) and phage λ transition between lysogenic and lytic growth (Savageau and Fasani, 2009)]. However, in each of these examples, the construction and analysis of the system’s design space was significantly simplified by automating and systematizing the System Design Space method. This was first

made possible via a collection of software tools known as the Design Space Toolbox for Matlab® (Fasani and Savageau, 2010).

The Design Space Toolbox for Matlab® provided a series of innovations that systematized the analysis of complex systems: it automated (1) construction of a System Design Space; (2) enumeration of the qualitatively-distinct phenotypes of a given system; and (3) the local analyses of the dominant S-System equations. Through these innovations it has provided insight into the fundamental principles underlying a variety of natural systems (Savageau and Fasani, 2009; Tolla and Savageau, 2011). Although these tools have paved the way for more complicated systems to be analyzed by the System Design Space method, it was clear that the implementation of these tools had severe limitations as it pertains to performance when analyzing larger systems. Here we present a second iteration of software tools, the Design Space Toolbox V2, that redesigns the computational approach, enables more complex circuitry to be analyzed, and extends the possible analyses through additional functionality.

Technology Overview

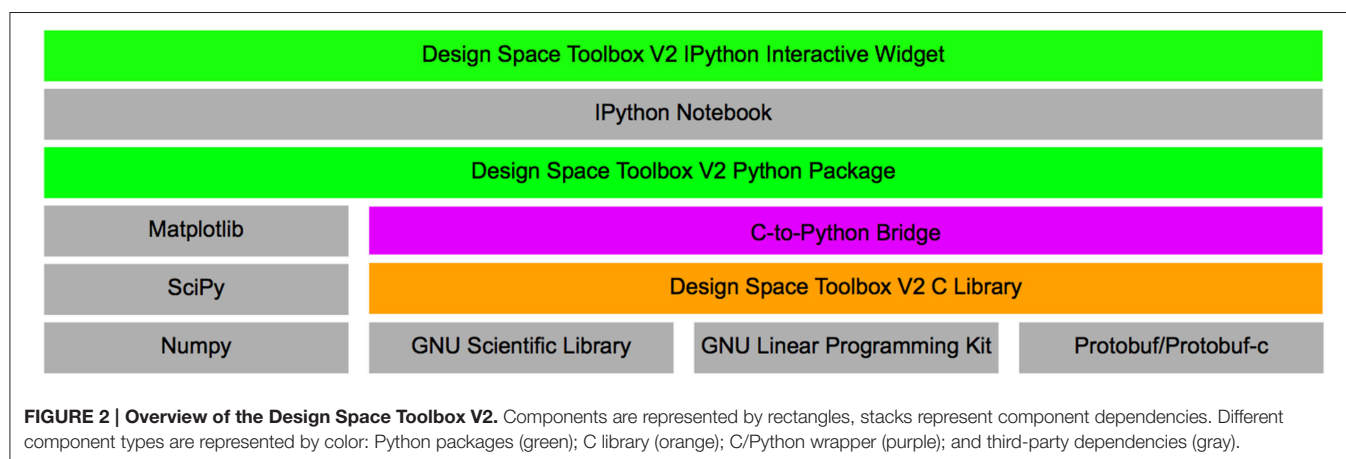
The original toolbox was built within Matlab® as a collection of .m scripts. There were many advantages that resulted from this decision: The Matlab® environment provided access to a variety of tools for symbolic algebra, linear algebra and linear optimization. Furthermore, it provides a rich scientific programming platform with its own interpreted language for rapid iterations between model formulation and model analysis. Furthermore, it provides fast vectorized operations that performed much better than iterated loops in its own language. These properties of the environment were critical in the design choices for the original toolbox, which improved performance by applying vector operations where possible and by providing an application programming interface that was part of the larger Matlab® ecosystem.

However, with these design choices come several limitations: The Matlab® environment provides access to limited system resources and its use of vectorized operations for faster performance had huge memory requirements that limited feasibility for larger problems.

Here, we present a novel set of tools using very different design choices. This new collection of tools is comprised of a stand-alone library, written in the C language, that implements its own symbolic algebra engine and leverages open-source compiled libraries for linear algebra (Gough, 2009) and linear optimization (via the GLPK library). This new toolbox applies concurrent approaches to leverage the “embarrassingly parallelizable” nature of the System Design Space approach by analyzing each qualitatively-distinct phenotype of the system independently from every other qualitatively-distinct phenotype using multi-threaded concurrent algorithms. A visual representation of the technology in the Design Space Toolbox V2 is shown in **Figure 2**.

This new software library applies many of the same concepts and theory of the previous version to automate the construction of a system design space, but involves a complete redesign of the tools for better memory management and parallelization for concurrent analysis. It also extends the original toolbox by providing an extensive library, with over 648 exposed functions, for the analysis of the system and its phenotypes. The new functionality of the toolbox includes: (1) automating the local stability analysis for model phenotypes; (2) enumerating the vertices of the feasible regions in up to three dimensions, both numerically and symbolically; (3) extending the capabilities of the symbolic algebra component to facilitate the analytical discovery of design principles; (4) defining constraints on the dependent variables and parameters of the system (i.e., to define architectural constraints and biological constraints), among many others.

The most important innovation provided by these new software tools is the enabling of a radically new modeling strategy. It does this by facilitating prediction of values for the parameters that can be used to focus computational effort on regions of parameter space that exhibit characteristics of particular interest. It achieves this automatically by (1) enumerating the repertoire of model phenotypes; (2) predicting sets of parameter values for any model phenotype; (3) predicting sets of parameter values for the simultaneous realization and visualization of an ensemble of model phenotypes; and (4) predicting sets of parameter values for the simultaneous realization of an ensemble of model phenotypes that are phased



to achieve a specific progression of behaviors (Lomnitz and Savageau, 2015).

Using this approach, we have been able to identify parameter values for a class of systems that display rich behaviors including monostability, bistability, sustained oscillations, and bifurcations among them (Lomnitz and Savageau, 2013, 2014, 2015).

Analysis via the Python Package

In this sub-section we illustrate the steps involved in constructing and analyzing a mechanistic model by an application of the toolbox to the simple example given by Equations (8) and (9). These methods will then be applied to a more complicated circuit that exhibits more interested behaviors in Section Example Applied to a Synthetic Memory Module. The simple example and the examples found later in this article illustrate the use of the Python Package within the Design Space Toolbox V2. This level of the toolbox offers access to most of the power of the C library within an interpreted environment similar to Matlab® for rapid scientific programming and prototype analyses.

The Design Space Toolbox V2 also includes a graphical user interface embedded within the IPython Notebook that facilitates its use by new users, with examples readily available online. However, the analyses presented in this Article are not reproducible using the graphical user interface and require the Python Package that has greater access to a wider set of functions.

Construction of the System Design Space

The first step in the analysis using the Design Space Toolbox V2 is to prepare the Python environment, which requires importing the python package into the current session:

```
import dspace
```

Once the environment has been initialized, the next step is to construct the appropriate computational objects that are used to formulate and analyze the mechanistic model. This entails refactoring the system of equations into a computer-readable format, simply a list of equations using a string representation (* represents multiplication, ^ represents power operator, d/dt). The differential equations and algebraic constraints are expressed explicitly by defining both sides of the equations. For example, the system described by Equations (8) and (9) is represented using the following string representation,

```
string_eq = ['X1. = a1*X100^-1\
+ a1*rho1*X1^2*K1^-2*X100^-1\
+ a1*X3*K3^-1*X100^-1\
- b1*X1\
- k*X1*X2',
'0 = 1 + X1^2*K1^-2 + X3*K3^-1 - X100'
]
```

and then the equations are parsed by the symbolic algebra component to construct an object of the `Equations` class that represents the system equations including auxiliary variables, which must be defined explicitly:

```
equations = dspace.Equations(string_eq,
                               auxiliary_variables=['X100'])
```

At this point, each string is used to construct an object of the `Expression` class, which parses the string and builds an abstract syntax tree representation that handles symbolic manipulation and evaluation of mathematical expressions within the design space toolbox. The `Equations` class can then be used to construct an object of the `DesignSpace` class,

```
ds = dspace.DesignSpace(string_eq)
```

that handles the majority of the steps involved in the System Design Space method. It calculates the maximum number of phenotypes, constructs objects that represent qualitatively-distinct phenotypes, and provides utility functions for visualization of the system design space. By convention, the `DesignSpace` object that represents the biochemical system being analyzed is named `ds`—a short name for convenience because it is the starting point for so many analyses.

Enumeration of the Phenotypic Repertoire

As mentioned in the previous sub-section, the `DesignSpace` object for a particular system is the starting point for most analyses. Among these analyses, perhaps the most important, is automatic enumeration of the phenotypic repertoire for a system. This is achieved by instructing the `ds` object to identify all its valid cases,

```
ph = ds.valid_cases()
```

where the output, stored in the “ph” variable, contains a list of case numbers for all the cases that have a feasible region somewhere in parameter space. These represent the qualitatively-distinct phenotypes of the system and together they define its phenotypic repertoire.

It should be noted that the `ds` object enumerates the phenotypic repertoire in parallel by creating a pool of cases that need to be tested, spawning n threads (where n is the number of processors available) that each request a case from the pool, and analyzing each case for validity. The results from each thread are returned to the `ds` object so it can assemble, sort and return the results. It should be noted that this is typically one of the most costly operations in an analysis and the results are stored by the `ds` object to eliminate excessive work following consecutive calls.

This can be applied to the example system and the number of valid phenotypes counted,

```
len(ph)
```

to show that it has a total of 10 qualitatively-distinct phenotypes that are valid somewhere in design space, as shown by the cases in **Table 1**.

TABLE 1 | Enumeration of the phenotypic repertoire for the simple system shown in Figure 1.

| Case number | Case signature | $\partial \log X_1 / \partial \log X_2$ | $\partial \log X_1 / \partial \log X_3$ | Stability |
|-------------|----------------|---|---|-----------|
| 1 | [1111] | 0.0 | 0.0 | S |
| 4 | [1211] | -1.0 | 0.0 | S |
| 7 | [2111] | 0.0 | 0.0 | U |
| 8 | [2121] | 0.0 | 0.0 | S |
| 9 | [2131] | 0.0 | 1.0 | U |
| 10 | [2211] | 1.0 | 0.0 | U |
| 11 | [2221] | -1.0 | 0.0 | S |
| 12 | [2231] | 1.0 | 1.0 | U |
| 15 | [3131] | 0.0 | 0.0 | S |
| 18 | [3231] | -1.0 | 0.0 | S |

Phenotypic Characteristics of Qualitatively-Distinct Phenotypes

The qualitatively-distinct phenotypes of the system can be analyzed for a variety of system properties, or phenotypic characteristics, such as those previously discussed in Section Mathematical Definition of Qualitatively-Distinct Phenotypes. Many phenotypic characteristics are automatically determined by the toolbox and these are typically determined by analyzing instances of the *Case* class that represent different cases of the design space. Instances of the *Case* class are obtained by calls to the *ds* object using a case identifier. For example, we can create a *Case* object representing Case 1 by calling the *ds* object with the case number 1 (or '1'),

```
case1 = ds(1)
```

or with the case signature [1111],

```
case1 = ds('1111', by_signature=True)
```

The phenotypic characteristics of a qualitatively-distinct phenotype typically fall within one of two categories: characteristics of the phenotype in the context of system design space (e.g., the boundaries of validity and global tolerance of the system to large qualitative changes; Coelho et al., 2009) or characteristics of the phenotype as they pertain to sub-system behavior (e.g., stability of the steady state and local robustness (insensitivity) of the system to small quantitative changes). In particular, our methods provide a novel means of characterizing global robustness, which we term “global tolerance” to clearly distinguish it from local robustness. Global tolerance is defined as the largest change in parameter values before there is a qualitative change in the phenotype (Coelho et al., 2009). This is determined automatically for each parameter and phenotype. Local robustness also is determined automatically by means of conventional parameter (in)sensitivities (local relative derivatives) for each parameter and phenotype. Some examples showing how this information is utilized in a stochastic context can be found in Fasani and Savageau (2013, 2015).

In general, characteristics in the context of system design space are determined from the *Case* object, and characteristics

in terms of sub-system behavior can be acquired from instances of the *SSystem* class that represent the dominant S-System of a particular case. The *SSystem* object representing the dominant S-System for a case is a property of the *Case* object. For example, the *SSystem* instance of Case 1 can be retrieved by

```
ssys = case1.ssystem
```

and the properties of the dominant S-System can be readily determined. For example, we can view the equations of the dominant S-System,

```
ssys.equations
```

which returns

```
[X1.=X100^-1*a1-X1*b1, X100=1]
```

a list of *Expression* objects represented using strings. Similarly, the steady-state solution for the *SSystem* object can also be viewed using the

```
ssys.solution
```

command that returns

```
[X1=a1*b1^-1, X100=1]
```

The *SSystem* class can be used to show (1) the sub-system equations, (2) analytical steady-state solution—in Cartesian and logarithmic coordinates, (3) numerical values for the steady-state solution at a given point, (4) the steady-state fluxes at a given point, (5) local factors like logarithmic gains for signal amplification and parameter sensitivities for local robustness, and (6) information concerning the local stability of the system.

In the last three columns of **Table 1** we show the logarithmic gains for X_1 with respect to X_2 and X_3 and the stability of a representative fixed point for a particular case. These two types of characteristics are acquired for the sub-system from the *SSystem* object. We begin by showing how an instance of the *SSystem* class can be analyzed for its phenotypic characteristics.

Automated Analysis of Log-Gain Factors and Parameter Sensitivities

The logarithmic gains and parameter sensitivities are purely a function of the kinetic orders of an S-System (Savageau, 2009). In the context of System Design space, the kinetic orders are architectural features of the system and thus for any particular system design are assumed to be constant (Lomnitz and Savageau, 2015); therefore, the log-gain factors and parameter sensitivities are constant for a particular dominant S-System. Furthermore, because independent variables and parameters of the system are treated equally, the parameter sensitivities are obtained in the same way as log-gain factors.

The logarithmic gain of X_1 relative to X_3 for Case 1 of the example is determined by

```
ssys.log_gain('X1', 'X3')
```

which, as shown in **Table 1**, is equal to 0—indicating that X_1 is uncoupled from X_3 and thus a change in X_3 does not elicit a change in X_1 as long as Case 1 is the qualitatively-distinct phenotype of interest.

Automated Analysis of Local Stability

As discussed previously, the `SSystem` class automates the determination of local stability for the S-Systems. However, for conventional eigenvalue analysis, the first step involves converting the dominant S-Systems into a purely dynamical system by removing any algebraic constraints. This is necessary because most dominant S-Systems include algebraic constraints originating from the recasting process. The algebraic constraints can typically be removed by virtue of the fact that S-System equations have tractable steady-state solutions; hence, the auxiliary variables can be solved in terms of the dynamic variables and system parameters, and their solution can then be used to eliminate the algebraic constraints. Here, we show how a solution for the auxiliary variables can be determined using the algebraic constraints to create a new representation of the dominant S-System that is purely dynamical.

Removing algebraic constraints from differential-algebraic S-Systems

We begin with the equations for a dominant S-System composed of ODEs and algebraic constraints, as shown in Equations (10) and (11). The algebraic constraints, shown in Equation (11), are equivalent to n_c equations of an S-System at steady state, where n_c is the number of auxiliary variables. The steady-state solution for the S-System equations can be readily obtained by transforming the system into logarithmic coordinates (Savageau, 2009) that are represented in matrix notation by the following equation

$$Gy + a = Hy + b \quad (16)$$

where G and H are $n_c \times (n + m)$ matrices of kinetic orders for the positive and negative terms of the algebraic equations, respectively—such that $G_{ij} = g_{ijp_i}$ and $H_{ij} = h_{ijq_i}$; y is an $(n + m)$ -column vector—such that $y_i = \log X_i$; a and b are n_c -column vectors of the logarithm of the rate constants for the positive and negative terms of the algebraic equations—such that $a_i = \log \alpha_{ip_i}$ and $b_i = \log \beta_{iq_i}$.

Next, we partition the G and H matrices into sub-matrices corresponding to dynamic, auxiliary and independent variables, represented by the t , c and I subscripts. Likewise, we partition the y vector into vectors corresponding to dynamic, auxiliary and independent variables,

$$y = \begin{bmatrix} y_t \\ y_c \\ y_I \end{bmatrix} \quad (17)$$

$$G = \begin{bmatrix} G_t & G_c & G_I \end{bmatrix} \quad (18)$$

$$H = \begin{bmatrix} H_t & H_c & H_I \end{bmatrix} \quad (19)$$

which yields the following system of equations in matrix notation,

$$G_t y_t + G_c y_c + G_I y_I + a = H_t y_t + H_c y_c + H_I y_I + b \quad (20)$$

We rearrange the terms so that the auxiliary variables are on the left-hand side of the equation and all other variables are on the right-hand side,

$$A_c y_c = -A_t y_t - H_I y_I + B \quad (21)$$

where $B = b - a$, and $A_i = G_i - H_i$ for $i = \{t, c, I\}$.

We find the inverse of A_c , defined as M_c , and multiply both sides of the equation, which yields the following equation for the auxiliary variables,

$$y_c = -R y_t - S y_I + U \quad (22)$$

where $R = M_c A_t$ is an $n_c \times n_t$ matrix; $S = M_c A_I$ is an $n_c \times m$ matrix; and $U = M_c B$ is an n_c -column vector. The solution for the i -th auxiliary variables, in Cartesian coordinates, is therefore

$$X_i = u_i \prod_{j=1}^{n_t} X_j^{r_{ij}} \prod_{j=n}^{n+m} X_j^{s_{ij}} \quad (23)$$

where u_i is the entry in the i -th row of the U vector; r_{ij} is the entry in the i -th row and $(j-n_t)$ -th column of the R matrix; and s_{ij} is the entry in the i -th row and $(j-n)$ -th column of the S matrix.

Substituting the solution for the auxiliary variables into the dynamic equations of the dominant S-System yields the following system of n_t ODEs,

$$\begin{aligned} \frac{dX_i}{dt} = & \alpha_{ip_i} \prod_{j=1}^{n_c} u_j^{g_{ijp_i}} \prod_{j=1}^{n_t} X_j^{g_{ijp_i} - \sum_{k=n_t}^n g_{ikp_i} r_{kj}} \prod_{j=n}^{n+m} X_j^{g_{ijp_i} - \sum_{k=n_t}^n g_{ikp_i} s_{kj}} \\ & - \beta_{iq_i} \prod_{j=1}^{n_c} u_j^{h_{ijq_i}} \prod_{j=1}^{n_t} X_j^{h_{ijq_i} - \sum_{k=n_t}^n h_{ikq_i} r_{kj}} \prod_{j=n}^{n+m} X_j^{h_{ijq_i} - \sum_{k=n_t}^n h_{ikq_i} s_{kj}} \end{aligned} \quad (24)$$

which no longer has algebraic constraints and thus can be analyzed using conventional S-System analysis for local stability (Savageau, 2009).

To remove algebraic constraints for an instance of the `SSystem` class, we use the following command,

```
alt_ssystem = ssys.remove_algebraic_constraints()
```

which creates a new instance of the `SSystem` class that is equivalent mathematically, but does not have the auxiliary variables and associated algebraic constraints. The equation of the `SSystem` object representing the dominant S-System for Case 1, without the algebraic constraints, is

```
[X1.=a1-X1*b1]
```

Analyzing purely dynamical S-Systems for local stability

The `SSystem` object, without algebraic constraints, is then analyzed for its local stability using one of two methods: standard eigenvalue analysis or by applying the Routh criteria for stability (Routh, 1877; Yang, 2002). In either case, the stability of a dynamical system depends on a particular set of values for the parameters.

Starting with a reference parameter set, stored in the `pvals` variable of the `VariablePool` class, that represents a point in design space for the S-System, we determine the eigenvalues for the system,


```
alt_ssystem.eigenvalues(pvals)
```

or we can quickly get the number of eigenvalues with positive real part,

```
alt_ssystem.positive_roots(pvals)
```

The stability of Case 1, at this representative point, is stable, as shown in **Table 1**, given that it has 0 eigenvalues with positive real part.

Automated Analysis of Global Tolerance

The Global Tolerance for a parameter and phenotype can be determined automatically from the `Case` object. From a geometric perspective, the global tolerance from a point is defined as the distance to the nearest boundaries of the enclosing phenotypic region in logarithmic coordinates (Fasani and Savageau, 2010). This can be calculated by performing a series of 1-D linear programming problems where all the parameters are fixed except for the parameter of interest.

The complete set of global tolerances for the example Case 1 at the starting reference parameter set is determined by

```
tolerances = Case1.measure_tolerances(pvals)
```

which returns a dictionary of key-value pairs, where the keys are the names of the parameters and the values are tuples with fold-decrease and fold-increase values representing the global tolerances in arithmetic coordinates, such that

```
tolerances['X3']
```

yields the tuple $(1e-20, 10.0)$. The first value indicates that a fold-decrease of 20 orders of magnitude are necessary to elicit a qualitative change in system behavior, whereas a 10 fold-increase results in a qualitative change in system behavior. Note that the value of $1e-20$ is in fact bounded by the program, and typically corresponds to an infinitely large global tolerance—hence a qualitative change in system behavior cannot be achieved by only decreasing the value of X_3 . Other large but fixed values may be determined by physical constraints such as the solubility limits for a metabolite or the diffusion limit for a particular rate constant. The set of global tolerances for Case 1, given a representative interior point, are shown in **Table 2**.

For this property, as for local stability in the previous section, starting from a representative point begs the question: How do we find this representative point?

Predicting Phenotype-Specific Parameter Sets

One of the challenges when analyzing non-linear systems is finding parameter values that realize a particular behavior or, in the context of the System Design Space method, a particular qualitatively-distinct phenotype of the system. For example, this mathematical model has a total of 12 independent variables/parameters that together define a 12-dimensional space. The naive approach might be to sample this space to try and find a combination that yields a particular phenotype of interest. However, even if we were to sample five values for each of the

TABLE 2 | Global tolerances for Case 1 of the system in Figure 1 measured as the fold-difference for a qualitative change in phenotype.

| Parameter | Global tolerance | |
|------------|---------------------|--------------------|
| | Fold-decrease | Fold-increase |
| α_1 | 1×10^{-20} | 3.162 |
| β_1 | 0.316 | 1×10^{20} |
| ρ_1 | 0.1 | 10.0 |
| K_1 | 0.316 | 1×10^{20} |
| K_3 | 0.1 | 1×10^{20} |
| k | 1×10^{-20} | 10.0 |
| X_2 | 1×10^{-20} | 10.0 |
| X_3 | 1×10^{-20} | 10.0 |

12 parameters, the number of combinations we would have to test would be enormous— $5^{12} = 244,140,625$; thus, this approach to search for values that realize a phenotype of the system is not feasible for most biological systems that have many more independent variables/parameters.

We have recently developed methods within the framework of the System Design Space approach that automatically predict representative values for any phenotype of the system (Lomnitz and Savageau, 2015). This is automated by the Design Space Toolbox V2 using linear programming techniques that can quickly and efficiently find the solution for the optimization of a linear function within a feasible region delimited by linear bounds (Vanderbei, 1996). Our software tools predict a set of values for the parameters of Case 1 using a simple instruction,

```
pvals = case1.valid_parameter_set()
```

that results in a parameter set near a vertex of the feasible region. Alternatively, parameter sets within the interior of the feasible region of a phenotype can be obtained by a variety of methods (e.g., see Lomnitz and Savageau, 2015) and is done using the following command:

```
pvals = case1.valid_interior_parameter_set()
```

The results for the local stability of the qualitatively distinct phenotypes shown in the last column of **Table 1** were determined by predicting a set of parameter values in the interior and calculating the number of eigenvalues with positive real part. The particular parameter set predicted for Case 1 is: $K_1 = 1.00$; $K_3 = 10.00$; $X_2 = 1.00$; $X_3 = 1.00$; $\alpha_1 = 1.00$; $\beta_1 = 10.00$; $\rho_1 = 10.00$.

The possible sets of values that our tools can predict are effectively limitless. To focus the choices, we can (1) impose power law constraints on the dependent and independent variables of the system, (2) optimize a power law objective function, and (3) impose bounds on the permissible values for each of the parameters and independent variables. Each of these options is a simple command, e.g.,

```
case1 = ds(1, constraints=['X1 > 100'])
```

```
pvals = case1.valid_parameter_set
    (p_bounds={ 'X3':[1e-3, 1e3]},
    optimize='X1^2*X2^2*X3^-2'
    )
```

Predicting Ensemble-Specific Parameter Sets

We have previously developed methods that enable the prediction of parameter values for the simultaneous realization of an ensemble of model phenotypes (Lomnitz and Savageau, 2015). The types of ensembles for which these tools can predict a corresponding set of parameter values fall within three categories: (1) intersections of phenotypes at a single point in design space; (2) co-localization of phenotypes within a slice of design space; and (3) arrangement of phenotypes phased within a slice of design space to exhibit a particular progression of behaviors.

Predicting Parameter Sets for Case Intersections

The validity of the intersection of multiple cases in design space can be readily determined using linear programming methods (Fasani and Savageau, 2010). This is achieved by combining the N_i boundary conditions of n different qualitatively-distinct phenotypes (Fasani and Savageau, 2010). This is particularly useful to determine if a system can exhibit multi-stability such as bistable regimes for hysteretic switches. The Design Space Toolbox V2 extends the analysis of intersecting cases beyond determining their validity. It enables prediction of values for the parameters that yield such an intersection. We begin by defining an object of the `CaseIntersection` class, that inherits many of its properties from the `Case` class.

Using the example `DesignSpace` object defined in Construction of the System Design Space, with the phenotypic repertoire shown in **Table 1**, we can determine if the intersection of different ensembles of cases are mathematically possible and, if so, predict values for the parameters that lead to their realization. To illustrate the `CaseIntersection` class, we apply it to identify intersections of three phenotypes consistent with bistable regimes. We choose the first two stable cases, Cases 1 and 4, and the first unstable case, Case 7, as shown in the last column of **Table 1**,

```
case1, case4, case7 = ds([1, 4, 7])

in = dspace.CaseIntersection([case1, case4,
    case7])
```

With the `CaseIntersection` object, we can determine validity as if it were a `Case` object by using the following command,

```
in.is_valid()
```

which yields `False`. This indicates that the intersection of Cases 1, 4, and 7 does not exist - regardless of values for the parameters. We can select an alternative intersection of three cases by selecting the next possible stable case, Case 8, instead of Case 4,

```
case8 = ds(8)
```

```
alt = dspace.CaseIntersection([case1,
    case7, case8])
```

and we determine its validity,

```
alt.is_valid()
```

which yields `True`. This indicates that the intersection of Cases 1, 7, and 8 exists and we can now proceed to predict a set of values for the parameters that results in the realization of this intersection. As with the `Case` object, we can predict a set of values using the `valid_parameter_set` method or `valid_interior_parameter_set` method,

```
pvals = alt.valid_interior_parameter_set()
```

which yields the following set of values for the parameters: $K_1 = 1.00$; $K_3 = 10.00$; $X_2 = 1.00$; $X_3 = 1.00$; $\alpha_1 = 0.32$; $\beta_1 = 10.00$; $\rho_1 = 100.00$.

Predicting Parameter Sets for Case Co-Localizations

An extension of the `Case Intersection` concept is `Case Co-localization`. This concept involves identifying an ensemble of n phenotypes that are simultaneously realized, hence they are valid, within a *slice* of design space for which a given number of parameters or independent variables are allowed to change. These variables, known as slice variables, define an s dimensional slice through design space, where s is the number of slice variables (Lomnitz and Savageau, 2015). The qualitatively distinct phenotypes, Cases as we have defined them, are the phenotypes associated with parameter values located within a particular polytope in system design space. Such polytopes may abut one another, or they might be completely separated; the situation is difficult to visualize in a high-dimensional space. The objective of the case co-localization function is to find a set of parameter values, if it exists, that yields a “slice” through the high-dimensional space that allows simultaneous visualization of selected polytopes. An intuitive example would be to determine if two phenotypes, e.g., a wild type and diseased phenotype, are simultaneously realized and the transition visualized within a 2D slice, where one axis represents a genotypically determined parameter and the other an environmentally determined variable.

We have previously shown that the validity of case co-localizations can be determined without sampling parameter space, it can be used for an arbitrary number of phenotypes and can be done in an arbitrary number of dimensions (Lomnitz and Savageau, 2015). We begin by duplicating and renaming the slice variables for each case in the ensemble and combining the boundaries for each case with the newly defined variables (Lomnitz and Savageau, 2015). The result is an $(m + n(s - 1))$ -dimensional convex polytope in logarithmic space, where s is the number of slice variables that can be analyzed in the same way as the feasible regions for Cases and Case Intersections.

The `CaseColocalization` class inherits properties from the `CaseIntersection` class and can be analyzed in the same way as a `CaseIntersection` object. As an example, we define an ensemble of phenotypes, composed of cases 8, 12, 15, and 18

and apply methods similar to those described in the previous subsection. This ensemble for co-localization, with X_2 as the slice variable, is created as follows:

```
c8,c12,c15,c18 = ds([8, 12, 15, 18])

co = dspace.CaseColocalization([c8, c12,
                                c15, c18], ['X2'])
```

With the `CaseColocalization` object, we can determine validity of the ensemble by using the following command:

```
co.is_valid()
```

In this example it yields `True`. This indicates that there is a simultaneous realization of these behaviors within a slice of parameter space, and that there are sets of values for the parameters capable of realizing this ensemble. Moreover, the sets can be determined automatically by using the `valid_parameter_set` method or `valid_interior_parameter_set` method,

```
co.valid_interior_parameter_set()
```

which yields the following sets of values for the parameters: $K_1 = 1.00$; $K_3 = 0.10$; $X_3 = 1.00$; $\alpha_1 = 0.10$; $\beta_1 = 10.00$; $\rho_1 = 10000.00$; $X_{2,8} = 1.00$; $X_{2,12} = 100.00$; $X_{2,15} = 1.00$; $X_{2,18} = 100.00$ —where $X_{2,i}$ represents the value for the X_2 variable within the feasible region for Case i .

Predicting Parameter Sets for Specific Arrangements of Cases

The method of case co-localization determines if an ensemble of cases can be simultaneously realized within some s -dimensional slice of parameter space, automatically and independent of sampling this infinitely large space. However, it does not yield any information about how the cases in the ensemble are located in the s -dimensional slice relative to each other or other important landmarks in the system design space. However, because these co-localizations are extensions of the methods that analyze cases in design space, we can apply the same methods. In particular, recall that the validity of cases in design space can be determined within particular constraints, as shown briefly at the end of Predicting Phenotype-Specific Parameter Sets. These same methods can be applied to objects of the `CaseColocalization` class to achieve a particular progression of behaviors (Lomnitz and Savageau, 2015), by imposing a set of power law constraints among all variables including replicated slice variables.

To illustrate this, we will create an ensemble of cases 8, 12, 15, and 18 arranged in ascending numerical order, such that $X_{2,8} < X_{2,12} < X_{2,15} < X_{2,18}$, are located from left to right in the design space of the system. We can determine whether this arrangement is possible somewhere in parameter space and predict values for the parameters that yield this particular arrangement.

An arrangement is created in the Design Space Toolbox V2 by introducing a new co-localization and adding constraints between the replicated slice variables representing X_2 when

defining the co-localization. The replicated slice variables representing X_2 are defined with special notation using the following format: `<slice variable>_<index in colocalization>`. In this example, the slice variable is X_2 and the indices in the co-localization for cases 8, 12, 15, and 18 are 0, 1, 2, and 3, respectively. Thus, the arrangement is constructed by

```
c8,c12,c15,c18 = ds([8, 12, 15, 18])

arr = dspace.CaseColocalization
      ([c8, c12, c15, c18],

      ['X2'],

      constraints = ['$X2_0 < $X2_1',
                    '$X2_1 < $X2_2',
                    '$X2_2 < $X2_3',

                    ])
```

The arrangement is simply a Case Co-localization plus additional constraints; thus, we determine validity and predict values for the parameters in the same way as we did for co-localization,

```
arr.is_valid()
```

which yields `False`. The result is that the specific arrangement that we specified is not possible, regardless of values for the parameters. Using the same cases, we can try different relative arrangements and additional constraints as long as both sides of the inequality defining the constraints are power laws. As an example, consider another arrangement involving Cases 8, 12, 15, and 18 with a different order by flipping the sign for one of the inequalities,

```
c8,c12,c15,c18 = ds([8, 12, 15, 18])

arr = dspace.CaseColocalization
      ([c8, c12, c15, c18],

      ['X2'],

      constraints = ['$X2_0 < $X2_1',
                    '$X2_1 > $X2_2',
                    '$X2_2 < $X2_3',
                    '$X2_1 > $X2_3']

      )
```

The validity of this co-localization yields `True` and thus, we can predict a set of values that realizes this arrangement: $K_1 = 1.00$; $K_3 = 0.10$; $X_3 = 1.00$; $\alpha_1 = 0.01$; $\beta_1 = 1.00$; $\rho_1 = 10000.00$; $X_{2,8} = 0.10$; $X_{2,12} = 10.00$; $X_{2,15} = 0.10$; $X_{2,18} = 10.00$.

Visualizing the Design Space of Biochemical Systems

One of the powerful features of the System Design Space method is that it partitions parameter space into regions and this structure reveals breakpoints in the characteristics of the system. This structured space, the design space of a system, can readily be visualized to gain insight into the landscape of possible phenotypes that a system can exhibit. These software tools enable visualization through the `matplotlib` package, the `NumPy` package and the `SciPy` package (Oliphant, 2007; Millman and Aivazis, 2011; van der Walt et al., 2011). We illustrate the Design Space Toolbox V2 visualization tools by applying them to the example system using the parameter set from the intersection of Cases 1, 7, and 8. In addition, we show the visualization of the stability showing the number of eigenvalues with positive real part.

The first step is to import the `matplotlib` plotting package into the python environment,

```
from matplotlib.pyplot import *
```

and import the plotting extension for the `dspace` classes,

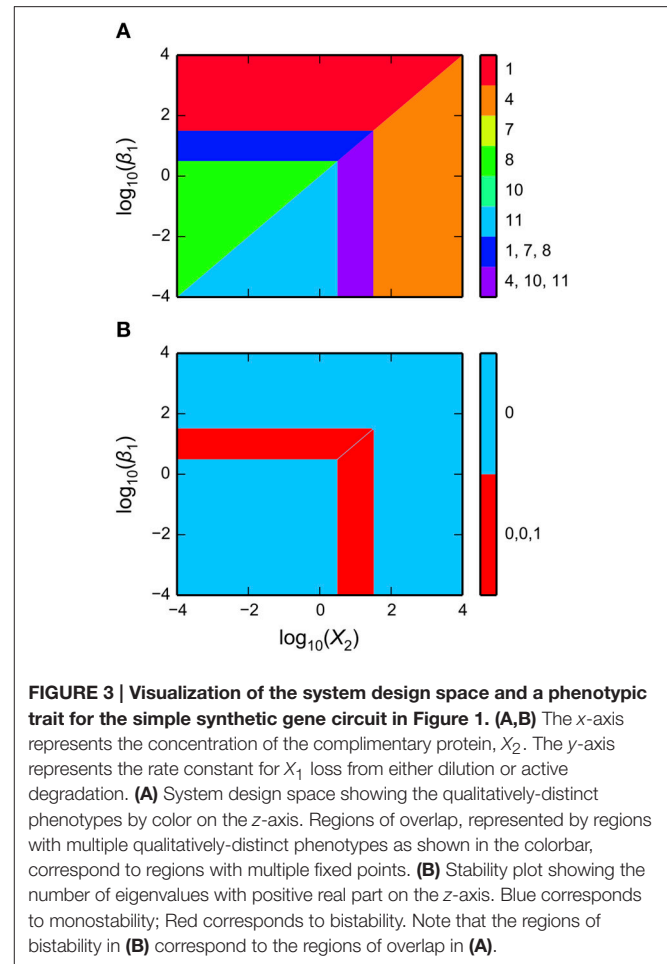
```
import dspace.plotutils
```

The typical way of visualizing a design space is by showing the qualitatively-distinct phenotypes in a 2D plot, where the x - and y -axes represent slice variables and the z -axis represents different cases identified by different colored regions.

Using the example `DesignSpace` object from the previous sub-sections, and the parameters predicted for the intersection of Cases 1, 7, and 8, these tools create the plot of the 2-D slice by the command

```
ds.draw_2D_slice(gca(), #:1
                pvals, #:2
                'X2', #:3
                'b1', #:4
                [1e-3, 1e3], #:5
                [1e-3, 1e3], #:6
                intersections = [1, 3] #:7
)
```

as shown in **Figure 3A**. The first argument is a `matplotlib` axis object for a plot canvas; the second argument is an instance of the `VariablePool` class with the values for the parameters; the third is the name of the x -axis; the fourth argument is the name of the y -axis; the fifth argument is the range of the x -axis in Cartesian coordinates; the sixth argument is the range of the y -axis in Cartesian coordinates; the seventh argument indicates the number of intersections



of cases to be drawn, where `[1, 3]` indicates it will display regions associated with individual phenotypes and with three phenotypes consistent with bi-stability (i.e., 2 stable and 1 unstable).

The stability of the fixed points also can be visualized. This is achieved by using a different command, but with mostly the same set of arguments,

```
ds.draw_2D_positive_roots(gca(), #:1
                          pvals, #:2
                          'X2', #:3
                          'b1', #:4
                          [1e-3, 1e3], #:5
                          [1e-3, 1e3], #:6
)
```

as shown in **Figure 3B**.

The toolbox provides additional tools to visualize dominant eigenvalues, steady-state concentrations, steady-state fluxes, and

mathematical functions evaluated at steady state. It also provides tools to visualize 1-D slices and 1-D response curves including stability information for bifurcation plots.

EXAMPLE APPLIED TO A SYNTHETIC MEMORY MODULE

In this Section, we illustrate the general capabilities of the Design Space Toolbox V2 by applying it to a two-gene synthetic circuit involving two transcriptional activators. This example serves the dual purpose of highlighting the novel, phenotype-centric, modeling strategy we have recently developed that inverts many of the typical steps in the conventional, parameter-centric, modeling approach (Lomnitz and Savageau, 2015).

The novel modeling approach begins by enumerating the phenotypic repertoire for a global perspective of system behavior; then predicting phenotype-specific or ensemble-specific parameter sets that realize phenotypic characteristics of interest; and finally focusing computational effort on localized regions of the parameter space for detailed analysis of the full system. The Design Space Toolbox V2 enables this novel modeling approach by automating the most difficult steps in the process.

The synthetic gene circuit, proposed and analyzed here, is intended to serve as a genetic hysteretic switch that can exhibit multistability. We show that this circuit can “count” between three distinct states in a positive direction that increases the counter and in a negative direction that decreases the counter. We show that by coupling the circuit with a target gene, a reporter, it can transition between three distinct intensity levels in a step-wise manner.

In the following sub-sections we (1) describe the design of the synthetic gene circuit; (2) formulate a mathematical model that captures the mechanistic details of the interactions; (3) analyze the system using our phenotype-centric modeling strategy; and (4) show examples of instances of the system at predicted points in the system’s design space that exhibit a variety of behaviors.

Synthetic Gene Circuit Design

Synthetic gene circuits have been constructed to serve a variety of purposes (Lu et al., 2009). One prominent use for synthetic biology is to forward engineer biological systems to gain insight into fundamental design principles (Mukherji and van Oudenaarden, 2009). Some examples that apply principles from engineering to biological systems include rationally designed synthetic oscillators (Elowitz and Leibler, 2000; Atkinson et al., 2003; Tigges et al., 2009) and bistable switches (Gardner et al., 2000; Atkinson et al., 2003).

We apply similar principles for the design of a system with the potential to exhibit multistability. This implies that there are instances of the system that have multiple stable fixed points, also known as steady-state *attractors*, with an associated set of initial conditions that define the *basin of attraction* within which the system gravitates toward a particular fixed point in state space.

The design of the synthetic gene circuit, represented in **Figure 4**, is composed of two transcriptional activators, X_1 and

X_2 that autogenously control expression of their own genes; the result is two seemingly independent positive feedback loops. The X_1 and X_2 regulators are translationally fused with a dimerization domain that causes X_1 monomers to form heterodimers with X_2 monomers. The X_1 – X_2 dimers are inactive and targeted for degradation by cellular proteases, which results in a strong thermodynamic potential that makes heterodimer formation essentially irreversible. Transcription of the activator genes is repressed by a third regulator, X_3 , that binds to the upstream region of the gene for both X_1 and X_2 , sterically hindering the auto-activation. The role of this repressor in the system is to tune the behavior of the system. A cartoon of the proposed construct is shown in **Figure 4A**, and an abstraction of the gene circuit with the key interactions is shown in **Figure 4C**.

Mathematical Model

We formulate a mathematical model composed of ODEs for the synthetic gene circuit design in **Figure 4**. Given that there is a fast turnover of mRNA relative to protein, we assume that synthesis of protein directly tracks mRNA expression. Thus, we model modulation of transcription as having a direct effect on the rate of protein synthesis. The mathematical model is described by the following system of non-linear equations,

$$\frac{dX_1}{dt} = \alpha_1 \left[\frac{1 + \rho_1 \left(\frac{X_1}{K_1} \right)^2 + \frac{X_3}{K_3}}{1 + \left(\frac{X_1}{K_1} \right)^2 + \frac{X_3}{K_3}} \right] - \beta_1 X_1 - k X_1 X_2 \quad (25)$$

$$\frac{dX_2}{dt} = \alpha_2 \left[\frac{1 + \rho_2 \left(\frac{X_2}{K_2} \right)^2 + \frac{X_3}{K_3}}{1 + \left(\frac{X_2}{K_2} \right)^2 + \frac{X_3}{K_3}} \right] - \beta_2 X_2 - k X_1 X_2 \quad (26)$$

where α_i represents the basal level of expression for the synthesis of the i -th regulator; β_i represents the rate constant for loss of the i -th regulator by dilution due to exponential growth; ρ_i represents the capacity for activation by the i -th regulator; K_i represents the concentration of the i -th regulator for half-maximal regulation; and k represents the rate constant for X_1 – X_2 heterodimer formation.

Recasting Equations into the Generic GMA Form

We recast the mathematical model into the generic GMA form using the 5-step approach outlined in Recasting Equations into a Generic Form, which yields the following system of differential-algebraic equations,

$$\frac{dX_1}{dt} = \alpha_1 X_{100}^{-1} + \alpha_1 \rho_1 X_1^2 K_1^{-2} X_{100}^{-1} + \alpha_1 X_3 K_3^{-1} X_{100}^{-1} - \beta_1 X_1 - k X_1 X_2 \quad (27)$$

$$\frac{dX_2}{dt} = \alpha_2 X_{200}^{-1} + \alpha_2 \rho_2 X_2^2 K_2^{-2} X_{200}^{-1} + \alpha_2 X_3 K_3^{-1} X_{200}^{-1} - \beta_2 X_2 - k X_1 X_2 \quad (28)$$

$$0 = 1 + X_1^2 K_1^{-2} + X_3 K_3^{-1} - X_{100} \quad (29)$$

$$0 = 1 + X_2^2 K_2^{-2} + X_3 K_3^{-1} - X_{200} \quad (30)$$

where X_{100} and X_{200} are the auxiliary variables defined for the denominators in Equations (25) and (26), respectively. These

equations are then used as input to the Design Space Toolbox V2 for analysis as described in Section Construction of the System Design Space.

Computer-Aided Novel Modeling Strategy

We analyze the system using the phenotype-centric modeling strategy (Lomnitz and Savageau, 2015) that involves (1) establishing criteria for what constitutes the model phenotypes of interest, (2) enumerating the repertoire of model phenotypes, (3) identifying model phenotypes that exhibit the characteristics of interest, and (4) predicting values for the parameters that realize the desired behavior. We have previously used this strategy to identify phenotypes that exhibit the potential for oscillation (e.g., see Lomnitz and Savageau, 2015) or specific couplings between inputs and outputs to achieve binary logic functions (Lomnitz and Savageau, in press). Here, the phenotype-centric modeling strategy is applied to identify a variety of phenotypes including bistability, tristability and quadrastability.

Criteria for Model Phenotypes of Interest

The first step in the phenotype-centric modeling strategy is to establish criteria for what constitutes a phenotype of interest based on a set of phenotypic characteristics. Typical characteristics include the coupling between input and output, stability of the fixed points, quantitative local robustness to small

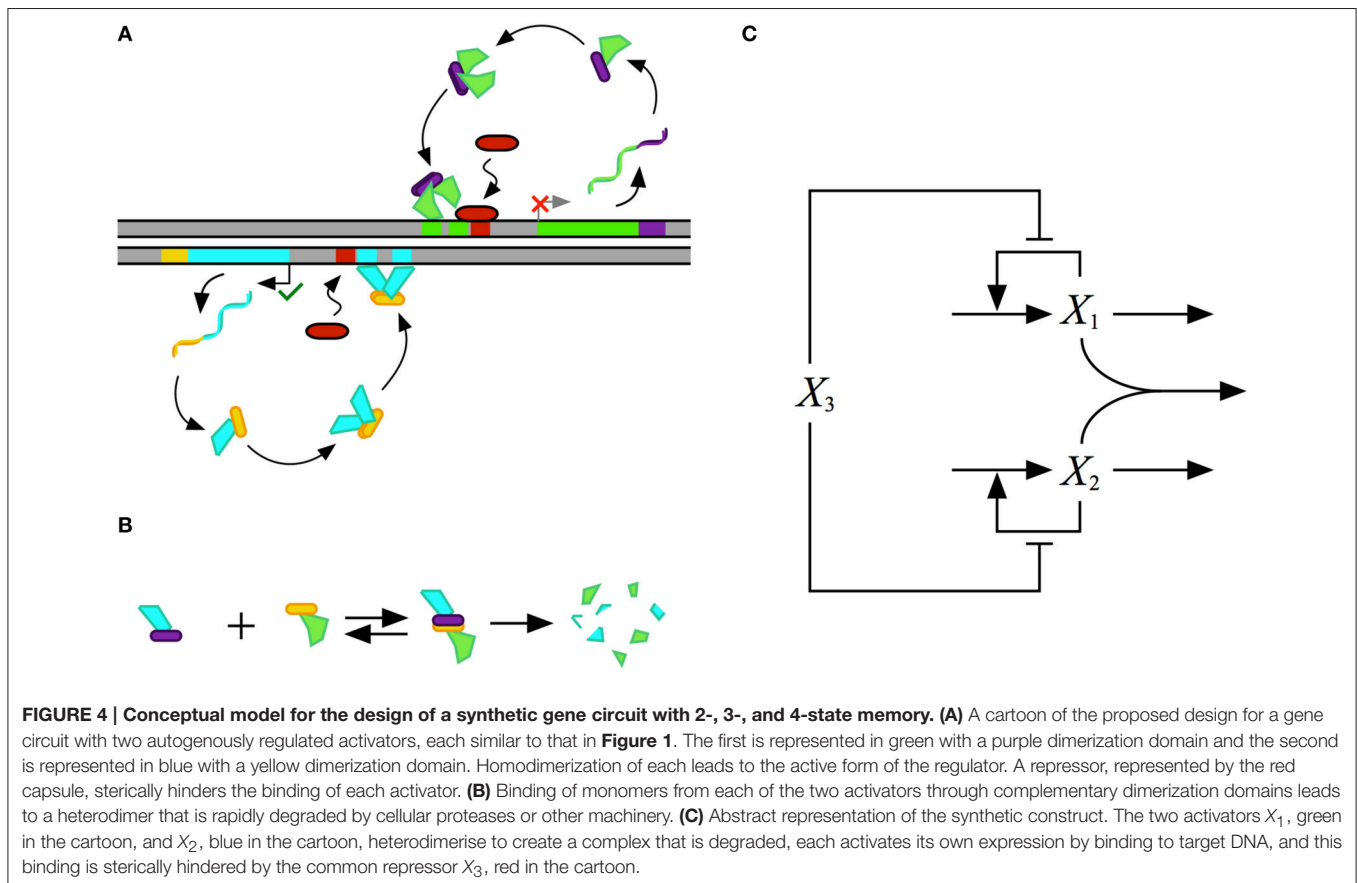
changes in system parameters, and qualitative global tolerance to large changes in system parameters.

The design for the synthetic gene circuit in **Figure 4** is expected to have the potential to exhibit multistability; therefore, there should be multiple fixed points, some of which are stable and some unstable, at a single point in parameter space. In the context of a system's design space, multistability involves an overlap or intersection of multiple cases (Savageau and Fasani, 2009; Fasani and Savageau, 2010; Martínez-Antonio et al., 2012).

Although multistability involves a combination of cases exhibiting either unstable or stable fixed points, we are interested in those that are stable; thus, the first criterion for what constitutes a phenotype of interest is that it be locally stable. Furthermore, a desirable property is that the fixed points be locally insensitive to unintended signals; thus, a second and third criterion is that both X_1 and X_2 are uncoupled from the repressor, X_3 . In summary, we are looking for cases that are locally stable, have X_1 uncoupled from X_3 [$L(X_1, X_3) = 0$], and have X_2 uncoupled from X_3 [$L(X_2, X_3) = 0$].

Enumerating the Repertoire of Phenotypes of Interest

The mechanistic model for the synthetic gene circuit is analyzed here following the outline in Section Design Space Toolbox V2: we (1) refactor the system equation into the



computer-readable format to construct a `DesignSpace` object, which we call `ds` (e.g., Section Construction of the System Design Space), (2) enumerate the valid phenotypes of the system using the `ds.valid_cases()` method (e.g., Section Enumeration of the Phenotypic Repertoire), and (3) determine the phenotypic characteristics of each valid phenotype to identify (a) the number of eigenvalues with positive real part at a representative point, (b) $L(X_1, X_3)$, and (c) $L(X_2, X_3)$ (e.g., Section Phenotypic Characteristics of Qualitatively-Distinct Phenotypes). The representative point to identify the number of eigenvalues with positive real part is predicted using the `valid_interior_parameter_set()` method of an instance of the `Case` class as described in Section Predicting Phenotype-Specific Parameter Sets. The number of phenotypes that satisfy our criteria are 21 of the 59 valid phenotypes, a portion of which is shown in **Table 3**.

Alternative Realizations of the Synthetic Gene Circuit

Maximizing the Number of Stable States

In Sections Predicting Phenotype-Specific Parameter Sets and Predicting Ensemble-Specific Parameter Sets we showed that our tools are able to predict values for the parameters that are specific to a phenotype or to an ensemble of phenotypes—either Case intersections at a single point in design space, Case co-localizations in a slice of design space, or Case specific arrangements in a slice of design space. Among these ensembles, Case intersections are particularly useful to identify the existence of multistability (Fasani and Savageau, 2010), and the ability of our tools to predict parameter values for their realization, as shown in Section Predicting Parameter Sets for Case Intersections, offers some interesting possibilities.

The first possibility we explore is the ability to identify the *maximum* number of stable phenotypes that can intersect in the system's design space, as this corresponds to the maximum number of steady-state attractors the system can exhibit. The general strategy on how to identify case intersections of n cases has been previously described (Fasani and Savageau, 2010). Here, we use this same approach but only apply it to the cases that are stable given that we are not interested in the cases that are unstable.

If the cases that satisfy the criteria are stored in the `cases` variable, our tools can list all the intersection of $k = \{2, 3, 4, \dots$,

$n\}$ cases. If for some value of k there are no intersections, the program stops and the value of $k-1$ is the maximum number of case intersections. The first step of finding all the intersections of $k = \{2, 3, 4, \dots, n\}$ cases is achieved by

```
attractors = ds.intersecting_cases
              (range(2, 22), cases)
```

and the result is a list of all possible intersections involving combinations of 2 up to 21 cases. These are stored in the `attractors` variable and used to identify the largest number of intersecting cases,

```
max([len(i._cases) for i in attractors])
```

which yields a maximum of four cases with stable fixed points that can be simultaneously realized at a single point in design space. Therefore, this design for a genetic memory module can have up to four steady state attractors for quadrastability.

Predicting Parameter Sets for Realization of Multi-Stability

The gene circuit design has a maximum of four steady-state attractors in which X_1 and X_2 can be high or low at any given time. This result might not be surprising, given that the system has two positive feedback loops that appear to be independent from each other. However, these positive feedback loops are part of an integrated system and can interact to produce interesting behaviors. One could speculate that an increase in either X_1 or X_2 might lead to a decrease in X_2 or X_1 , respectively, due to the formation of X_1 – X_2 heterodimers. Here, we explore a series of alternative behaviors for bistable, tristable and quadrastable switches including a stable counter with three different levels.

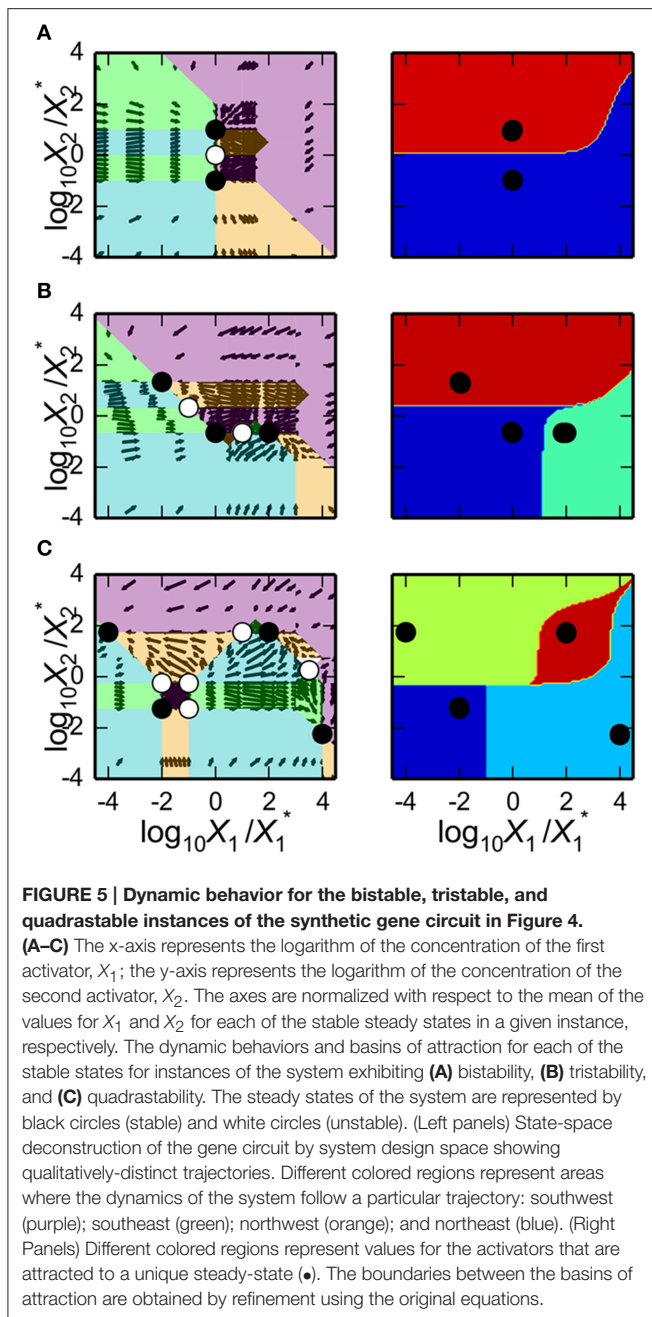
The System Design Space method we have described can be applied for a deconstruction of dynamic behaviors in state space. This deconstruction, which is still in the early phases of its development, partitions state space into regions that exhibit qualitatively-distinct trajectories that provide valuable information regarding the system's basins of attraction and response to transient perturbations. The dominance conditions define $(n + m)$ -dimensional polytopes, where n is the number of dependent variables and m is the number of independent variables/parameters. For each equation in the Dominant S-System, we can identify regions where the positive term is greater than the negative term and thus a region with a qualitatively-defined trajectory. The particular arrangement of steady states and the trajectories around these steady states can be represented visually, as shown in the left panels of **Figure 5**, and can be compared with the basins of attraction for the original system of equations, as shown in the right panels of **Figure 5**. In each of these examples, our automated tools provide rich information for rapid identification of interesting properties for the system. The results can then be refined by applying conventional methods to the full system.

Predicting bistable genetic switches

The Design Space Toolbox V2 can be used to predict values for the parameters that result in instances of the system that are only bistable switches. We achieve this in two steps: we identify all the

TABLE 3 | Enumeration of the phenotypic repertoire for the system shown in Figure 2.

| Case number | Case signature | $L(X_1, X_3)$ | $L(X_2, X_3)$ | Stability |
|-------------|----------------|---------------|---------------|-----------|
| 1 | 11111111 | 0.0 | 0.0 | S |
| 10 | 11121111 | 0.0 | 0.0 | S |
| 19 | 11211111 | 0.0 | 0.0 | U |
| ... | ... | ... | ... | ... |
| 297 | 32213131 | 1.0 | 1.0 | U |
| 306 | 32223131 | 0.5 | 0.5 | U |
| 315 | 32313131 | 0.0 | 0.0 | S |



valid ensembles of two stable phenotypes satisfying our criteria, and then predict representative parameter values and identify those instances that have only two steady-state attractors—to eliminate ensembles that might be part of higher-order ensembles with more steady-state attractors.

The first step is most easily achieved using the same command as in Section Alternative Realizations of the Synthetic Gene Circuit, modified to return only Case Intersections involving two stable phenotypes,

```
en2 = ds.intersecting_cases([2], cases)
```

where `en2` stores all the ensembles of two stable phenotypes at a single point.

The second step is achieved by iterating through each ensemble [`for en in en2;`]; predicting a representative point that realizes an ensemble [`pvals=en.valid_interior_parameter_set()`]; identifying the cases valid at the representative point [`all_cases = ds(ds.valid_cases(p_bounds=pvals))`]; and counting the number of cases that are locally stable [`sum([case.positive_roots() == 0 for case in all_cases])`]. An example from among the six showing a bistable instance of the design, as predicted following these steps, is shown **Figure 5A**.

Predicting tristable genetic switches

We identify instances of the system that exhibit tristability using the same approach used to identify bistability—we identify the valid ensembles of three stable phenotypes and select those that have only three steady-state attractors. We change the first step by identifying the ensembles with Case Intersections of three stable phenotypes,

```
en3 = ds.intersecting_cases([3], cases)
```

and proceed with the same steps used for the bistable case. We find eight ensembles that exhibit tristability, an example of which is shown in **Figure 5B**.

Predicting quadrastable genetic switches

Because the maximum number of stable phenotypes that can intersect at a given point in design space is 4, the task of identifying instances of the system that exhibit quadrastability is simpler than the bistable and tristable examples. Here, all we need to do is identify ensembles of four stable phenotypes,

```
en4 = ds.intersecting_cases([4], cases)
```

which yields a total of 18 that can exhibit quadrastability. An example is shown in **Figure 5C**.

Predicting State-Space Arrangements of the Steady-State Attractors

As we discussed in Section Predicting Phenotype-Specific Parameter Sets, we can add constraints to the system and thus the number of parameter sets we can predict is effectively limitless. Here, we show how constraints can be imposed to identify relative arrangements of the steady-state attractors that are permissible in state space. To achieve this, we define new independent variables that partition state space into four quadrants [i.e., $(-, -)$, $(-, +)$, $(+, -)$, and $(+, +)$] and apply our tools to determine which combination of quadrants the stable-state attractors can occupy.

We define two variables, $X_{r,1}$ and $X_{r,2}$, that partition state space into the four quadrants with the boundaries $X_1 = X_{r,1}$ and $X_2 = X_{r,2}$, such that the $(-, -)$ quadrant is given by $X_1 < X_{r,1}$ and $X_2 < X_{r,2}$. Then, we reconstruct a new instance of the `DesignSpace` class with the independent variables explicitly defined to include the $X_{r,1}$ and $X_{r,2}$ variables.

The new instance of the `DesignSpace` class can create new instances of the `Case` class with added constraints, as shown in Section Predicting Phenotype-Specific Parameter Sets.

From Section Predicting Parameter Sets for Realization of Multi-Stability we identified all the ensembles of four stable phenotypes that result in a quadrastable instance of the system. We then test the validity of each of these ensembles with constraints imposed on its constitutive cases. For example, if we have four cases with case identifiers represented by the variables `case0`, `case1`, `case2`, and `case3` that comprise an ensemble for a quadrastable system, we impose constraints on these cases to ensure that each is in a separate quadrant as follows,

```
C0 = ds(case0, constraints = ['X1 < Xr1',
                             'X2 < Xr2'])
```

```
C1 = ds(case1, constraints = ['X1 < Xr1',
                             'X2 > Xr2'])
```

```
C2 = ds(case2, constraints = ['X1 > Xr1',
                             'X2 < Xr2'])
```

```
C3 = ds(case3, constraints = ['X1 > Xr1',
                             'X2 > Xr2'])
```

```
ensemble = space.CaseIntersection([C0,
                                   C1, C2, C3])
```

and validity of the ensemble can be tested as shown in Section Predicting Ensemble-Specific Parameter Sets. We apply this to test each of the 35 combinations of criteria like that in the example above. We find that 24 of the 35 can satisfy their relevant criteria and that the remaining 11 are unable to satisfy their relevant criteria regardless of values for the parameters and thresholds for the quadrants.

Predicting a Stable Counter With Positive and Negative Channels

One arrangement of particular interest has one steady-state attractor that occupies each of the quadrants—consistent with four binary boolean states, represented by $(-, -)$, $(-, +)$, $(+, -)$, and $(+, +)$. We find that all of the ensembles identified in Section Predicting Parameter Sets for Realization of Multi-Stability are able to yield this particular arrangement of steady-state attractors, an example of which is shown in **Figure 6**, where $X_{r,1} = 1$ and $X_{r,2} = 1$.

This combination of $(-, -)$, $(+, -)$, $(-, +)$, and $(+, +)$ binary boolean states makes this design useful as a control switch where the expression of target genes are regulated by X_1 , X_2 or both. For example, this synthetic circuit, controlling a reporter gene whose synthesis is directly coupled to X_1 and inversely coupled to X_2 , can effectively count from 0 to 3 at well-defined levels for its expression. Such a reporter under the control of this module is modeled mathematically by the following ODE

$$\frac{dX_4}{dt} = \alpha_4 \left(\frac{\rho_{41}X_1^2 + K_1^2}{X_1^2 + K_1^2} \right) \left(\frac{\rho_{42}^{-1}X_2^2 + K_2^2}{X_2^2 + K_2^2} \right) - \beta_4X_4 \quad (31)$$

where X_4 represents concentration of the reporter protein; α_4 represents the rate of synthesis of X_4 at an unrepresed and inactivated state; β_1 represents the rate constant for loss of X_4 by

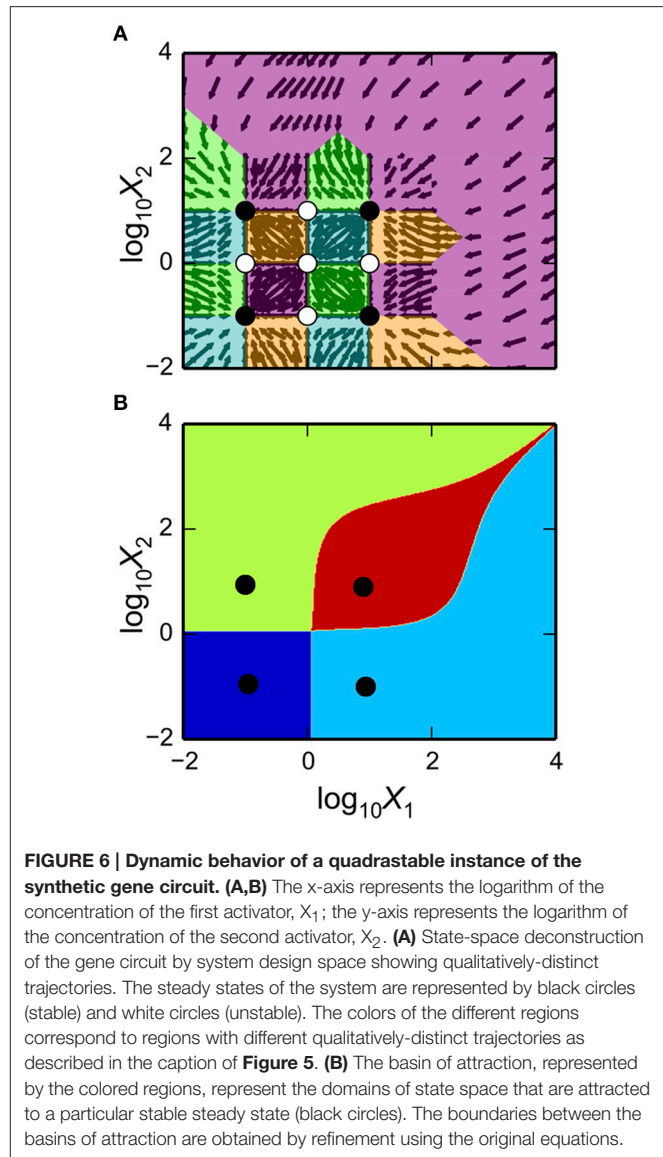


FIGURE 6 | Dynamic behavior of a quadrastable instance of the synthetic gene circuit. (A,B) The x-axis represents the logarithm of the concentration of the first activator, X_1 ; the y-axis represents the logarithm of the concentration of the second activator, X_2 . **(A)** State-space deconstruction of the gene circuit by system design space showing qualitatively-distinct trajectories. The steady states of the system are represented by black circles (stable) and white circles (unstable). The colors of the different regions correspond to regions with different qualitatively-distinct trajectories as described in the caption of **Figure 5**. **(B)** The basin of attraction, represented by the colored regions, represent the domains of state space that are attracted to a particular stable steady state (black circles). The boundaries between the basins of attraction are obtained by refinement using the original equations.

dilution due to exponential growth; ρ_{41} represents the capacity for activation of X_4 synthesis by X_1 ; ρ_{42} represents the capacity for repression of X_4 synthesis by X_2 .

The ability of this design to perform as a stable counter arises from the X_1 – X_2 heterodimer formation in combination with the seemingly independent positive feedback loops for X_1 and X_2 . For example, a transient increase in one species elicits a transient drop in the other that, in combination with the positive feedback loops, can lead to a switch from a stable “+” state to a stable “–” state.

This is reflected in the teardrop-shaped basin of attraction for the steady-state attractor in the $(+, +)$ quadrant: when the system is at the $(+, +)$ attractor and there is a transient increase in the concentration of either X_1 or X_2 , the dynamics of the system are such that it leaves the basin of attraction for the $(+, +)$ attractor and enters the basin of attraction for the $(+, -)$ or $(-, +)$ attractor, respectively. A visual representation of the

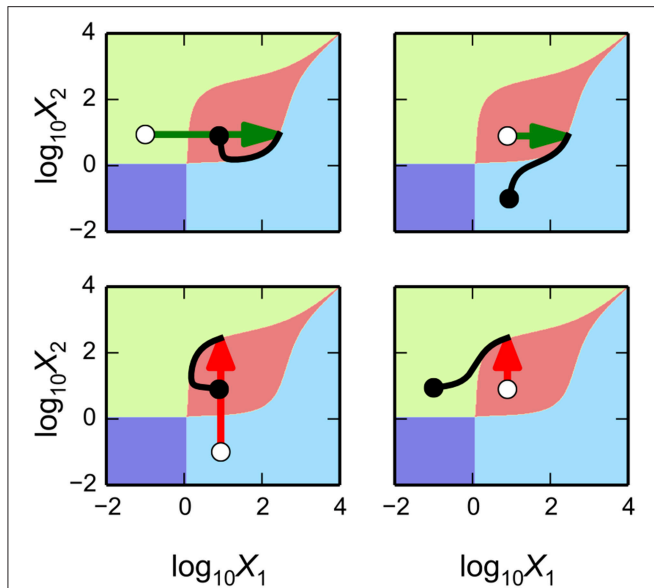


FIGURE 7 | Basins of attraction for a 4-state genetic counter. The x-axis represents the logarithm of the concentration of the first activator, X_1 ; the y-axis represents the logarithm of the concentration of the second activator, X_2 . Different colored regions represent values for the activators that converge to a unique steady-state attractor. Transitions from an initial steady state (white circle) to a new steady state (black circle) following an equal size bolus ($275 \mu\text{M}$) in one of the two activators. The top panels show transient simulations following a bolus of X_1 (green arrows). The bottom panels show transient simulations following a bolus of X_2 (red arrows). Left and right sub-panels show the transitions from different initial steady-state attractors.

transitions between the steady-state attractors following transient stimulation is shown in **Figure 7**.

Assume that the system is poised at the attractor in the $(-, +)$ quadrant; if X_1 is added in some amount, i.e., $275 \mu\text{M}$, the system transitions to the attractor in the $(+, +)$ quadrant; then if X_1 is added again in the same amount, a transition to the attractor in the $(+, -)$ quadrant ensues; therefore, by adding the same bolus of X_1 twice, in a step-wise fashion, the system has switched between an equal number of steps, which bears the signature of a genetic counter.

Now, assume the system is poised at the opposite attractor in the $(+, -)$ quadrant; if X_2 is added in the same amount the system transitions to the attractor in the $(+, +)$ quadrant; then if X_2 is added again in the same amount, a transition to the attractor in the $(+, -)$ quadrant ensues; therefore, by adding the same bolus of X_2 twice, in a step-wise fashion, the system has reverted back to the original state.

These traits show that the system has two distinct channels that enable two sequences of transitions between the same three states but in the opposite order. A positive channel for $(-, +) \rightarrow (+, +) \rightarrow (+, -)$ and a negative channel for $(+, -) \rightarrow (+, +) \rightarrow (-, +)$. By coupling the module with the reporter gene, we show that the system is capable of counting between three levels of reporter concentration and can perform basic arithmetic using values 0, 1, and 2. An example showing a sequence of

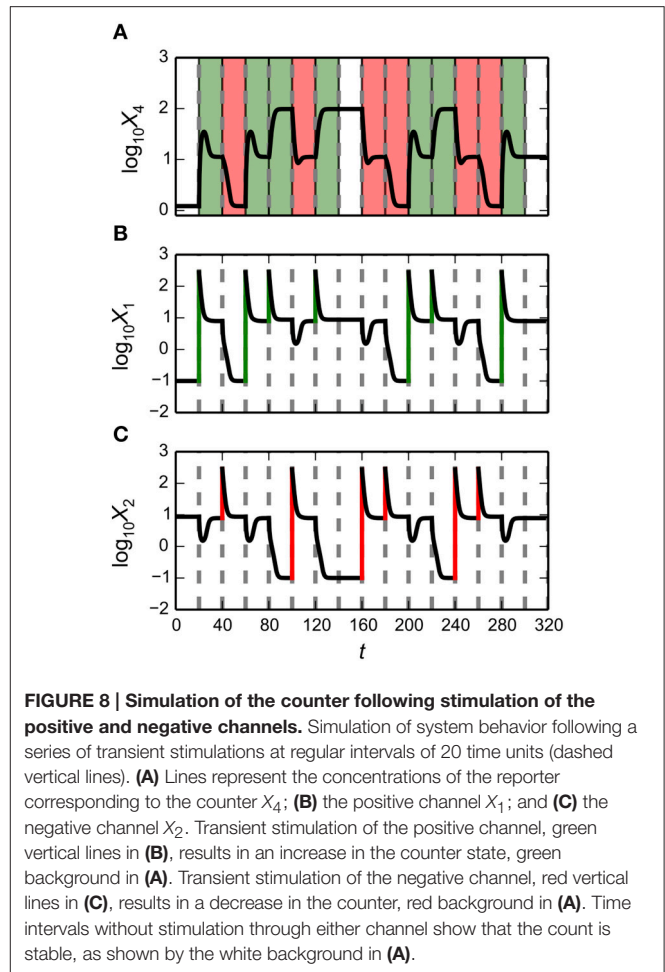


FIGURE 8 | Simulation of the counter following stimulation of the positive and negative channels. Simulation of system behavior following a series of transient stimulations at regular intervals of 20 time units (dashed vertical lines). **(A)** Lines represent the concentrations of the reporter corresponding to the counter X_4 ; **(B)** the positive channel X_1 ; and **(C)** the negative channel X_2 . Transient stimulation of the positive channel, green vertical lines in **(B)**, results in an increase in the counter state, green background in **(A)**. Transient stimulation of the negative channel, red vertical lines in **(C)**, results in a decrease in the counter, red background in **(A)**. Time intervals without stimulation through either channel show that the count is stable, as shown by the white background in **(A)**.

additions and subtractions following transient addition of X_1 and X_2 , respectively, is shown in **Figure 8**.

CONCLUSIONS

The Design Space Toolbox V2 is a compendium of tools designed to aid in the analysis and design of biochemical systems. It is particularly useful for the characterization of system design principles. Indeed, each of the “landmarks” in system design space—boundaries and vertices—are rigorously defined by particular constellations of parameter values that represent the “design principles” of the system (e.g., Savageau and Fasani, 2009). These constellations are not at all obvious and would be difficult to discover by trial and error, but are automatically determined with our tools. As in other engineering disciplines, knowing such design principles allows one to control the system in a more rational fashion.

These tools have already proven useful for understanding complex natural circuitry (Savageau, 2013) and for rationally designing and engineering new synthetic gene circuits (Lomnitz and Savageau, 2013, 2014, 2015) described by models composed of power functions from chemical kinetics and rational functions

from biochemical kinetics. However, the full scope of models that can be analyzed by these new tools has yet to be explored.

These tools automate the construction and analysis of the design space of biochemical systems in a manner similar to a previous iteration of software tools known as the design space toolbox for Matlab®. However, this new iteration is a complete redesign of the approach that expands the scope of applicable systems beyond what was previously possible due to limits on both time and computational resources. The most important contribution provided by these tools is the enabling of a radically new phenotype-centric modeling strategy (Lomnitz and Savageau, 2015) that inverts the steps in the conventional parameter-centric strategy and automates those that are most difficult.

To illustrate our software tools, we applied them to the design of a synthetic two-gene circuit that has positive feedback loops with the potential for hysteretic-switch behavior. However, unlike other hysteretic switch designs that exhibit typical bistability (e.g., Gardner et al., 2000; Atkinson et al., 2003), this circuit has two seemingly independent positive feedback loops that are coupled by a fused heterodimerization domain. In an automated analysis, we show that this design can be tuned to exhibit up to four stable steady states. Furthermore, our tools predict multiple sets of values for the parameters that realize specific instances of the system that exhibit bistability, tristability and quadrastability.

Further analysis of a quadrastable instance of the system reveals that it can alternate between three of the steady states following transient stimulation in one of two input channels: a positive channel that results in the forward transition between these states; and a negative channel that results in the reverse transition between these same states. By coupling this network

to a reporter gene, we have shown that this circuit can effectively count between three levels of fluorescence intensity in a step-wise manner.

These examples show the power of our new tools and illustrate how they enable a radically new modeling strategy that does not rely on first establishing nominal values for the parameters. Instead, this phenotype-centric strategy enumerates the phenotypic repertoire, identifies phenotypes of interest according to specific criteria, and then predicts sets of parameter values for realizing the phenotypes of interest. By assembling a variety of criteria, these tools can predict instances of a system that displays a rich assortment of behaviors.

AUTHOR CONTRIBUTIONS

MS conceived the initial approach; JL and MS developed the methodology and wrote the manuscript; JL developed the software.

FUNDING

This work was supported in part by a grant to MAS from the US Public Health Service, National Institutes of Health (RO1-GM30054).

ACKNOWLEDGMENTS

We thank Rick Fasani, Dean Tolla, and Pedro Coelho for their fruitful discussions regarding the System Design Space method and the Design Space Toolbox V2, and Alberto Marin-Sanguino and Christiana Sehr for helpful user feedback.

REFERENCES

- Atkinson, M. R., Savageau, M. A., Myers, J. T., and Ninfa, A. J. (2003). Development of genetic circuitry exhibiting toggle switch or oscillatory behavior in *Escherichia coli*. *Cell* 113, 597–607. doi: 10.1016/S0092-8674(03)00346-5
- Benner, S. A., and Sismour, A. M. (2005). Synthetic biology. *Nat. Rev. Genet.* 6, 533–543. doi: 10.1038/nrg1637
- Brenner, S. (2000). Genomics. The end of the beginning. *Science* 287, 2173–2174. doi: 10.1126/science.287.5461.2173
- Coelho, P. M. B. M., Salvador, A., and Savageau, M. A. (2009). Quantifying global tolerance of biochemical systems: design implications for moiety-transfer cycles. *PLoS Comput. Biol.* 5:e1000319. doi: 10.1371/journal.pcbi.1000319
- Elowitz, M. B., and Leibler, S. (2000). A synthetic oscillatory network of transcriptional regulators. *Nature* 403, 335–338. doi: 10.1038/35002125
- Fasani, R. A., and Savageau, M. A. (2010). Automated construction and analysis of the design space for biochemical systems. *Bioinformatics* 26, 2601–2609. doi: 10.1093/bioinformatics/btq479
- Fasani, R. A., and Savageau, M. A. (2013). Molecular mechanisms of multiple toxin-antitoxin systems are coordinated to govern the persister phenotype. *Proc. Natl. Acad. Sci. U.S.A.* 110, E2528–E2537. doi: 10.1073/pnas.1301023110
- Fasani, R. A., and Savageau, M. A. (2015). Unrelated toxin-antitoxin systems cooperate to induce persistence. *J. R. Soc. Interface* 12, 20150130. doi: 10.1098/rsif.2015.0130
- Friedland, A. E., Lu, T. K., Wang, X., Shi, D., Church, G., and Collins, J. J. (2009). Synthetic gene networks that count. *Science* 324, 1199–1202. doi: 10.1126/science.1172005
- Gardner, T. S., Cantor, C. R., and Collins, J. J. (2000). Construction of a genetic toggle switch in *Escherichia coli*. *Nature* 403, 339–342. doi: 10.1038/35002131
- Gough, B. (ed.). (2009). *GNU Scientific Library Reference Manual. 3rd Revised Edn.* Bristol: Network Theory Ltd.
- Kim, B., and Little, J. W. (1992). Dimerization of a specific DNA-binding protein on the DNA. *Science* 255, 203–206. doi: 10.1126/science.1553548
- Kim, J. K., and Forger, D. B. (2012). A mechanism for robust circadian timekeeping via stoichiometric balance. *Mol. Syst. Biol.* 8, 630. doi: 10.1038/msb.2012.62
- Lomnitz, J. G., and Savageau, M. A. (2013). Phenotypic deconstruction of gene circuitry. *Chaos* 23, 025108. doi: 10.1063/1.4809776
- Lomnitz, J. G., and Savageau, M. A. (2014). Strategy revealing phenotypic differences among synthetic oscillator designs. *ACS Synth. Biol.* 3, 686–701. doi: 10.1021/sb500236e
- Lomnitz, J. G., and Savageau, M. A. (2015). Elucidating the genotype-phenotype map by automatic enumeration and analysis of the phenotypic repertoire. *Npj. Syst. Biol. Appl.* 1, 15003. doi: 10.1038/npjbsa.2015.3
- Lomnitz, J. G. and Savageau M. A. (in press). Rapid discrimination among putative mechanistic models of biochemical systems. *Sri. Rep.*
- Lu, T. K., Khalil, A. S., and Collins, J. J. (2009). Next-generation synthetic gene networks. *Nat. Biotechnol.* 27, 1139–1150. doi: 10.1038/nbt.1591
- Mandal, N., Su, W., Haber, R., Adhya, S., and Echols, H. (1990). DNA looping in cellular repression of transcription of the galactose operon. *Genes Dev.* 4, 410–418. doi: 10.1101/gad.4.3.410
- Martínez-Antonio, A., Lomnitz, J. G., Sandoval, S., Aldana, M., and Savageau, M. A. (2012). Regulatory design governing progression of population growth phases in bacteria. *PLoS ONE* 7:e30654. doi: 10.1371/journal.pone.0030654

- Martin, V. J. J., Pitera, D. J., Withers, S. T., Newman, J. D., and Keasling, J. D. (2003). Engineering a mevalonate pathway in *Escherichia coli* for production of terpenoids. *Nat. Biotechnol.* 21, 796–802. doi: 10.1038/nbt833
- Millman, K. J., and Aivazis, M. (2011). Python for scientists and engineers. *Comput. Sci. Eng.* 13, 9–12. doi: 10.1109/MCSE.2011.36
- Monod, J., Wyman, J., and Changeux, J.-P. (1965). On the nature of allosteric transitions: a plausible model. *J. Mol. Biol.* 12, 88–118. doi: 10.1016/S0022-2836(65)80285-6
- Mukherji, S., and van Oudenaarden, A. (2009). Synthetic biology: understanding biological design from synthetic circuits. *Nat. Rev. Genet.* 10, 859–871. doi: 10.1038/nrg2697
- Oliphant, T. E. (2007). Python for scientific computing. *Comput. Sci. Eng.* 9, 10–20. doi: 10.1109/MCSE.2007.58
- Ro, D.-K., Paradise, E. M., Ouellet, M., Fisher, K. J., Newman, K. L., Ndungu, J. M., et al. (2006). Production of the antimalarial drug precursor artemisinic acid in engineered yeast. *Nature* 440, 940–943. doi: 10.1038/nature04640
- Routh, E. J. (1877). *A Treatise on the Stability of a Given State of Motion: Particularly Steady Motion*. London: Macmillan and Company.
- Savageau, M. A. (2009). *Biochemical Systems Analysis: A Study of Function and Design in Molecular Biology, 40th Anniversary Edn.* Charleston, SC: CreateSpace [A reprinting of the Original Edition, Reading, MA: Addison-Wesley Publishing; 1976].
- Savageau, M. A. (2013). “Phenotypes and design principles in system design space,” in *Handbook of Systems Biology: Concepts and Insights*, eds A. J. M. Walhout, M. Vidal, and J. Dekker (San Diego, CA: Elsevier), 287–310.
- Savageau, M. A., Coelho, P. M. B. M., Fasani, R. A., Tolla, D. A., and Salvador, A. (2009). Phenotypes and tolerances in the design space of biochemical systems. *Proc. Natl. Acad. Sci. U.S.A.* 106, 6435–6440. doi: 10.1073/pnas.0809869106
- Savageau, M. A., and Fasani, R. A. (2009). Qualitatively distinct phenotypes in the design space of biochemical systems. *FEBS Lett.* 583, 3914–3922. doi: 10.1016/j.febslet.2009.10.073
- Savageau, M. A., and Voit, E. O. (1987). Recasting nonlinear differential equations as S-systems: a canonical nonlinear form. *Math. Biosci.* 87, 83–115. doi: 10.1016/0025-5564(87)90035-6
- Stricker, J., Cookson, S., Bennett, M. R., Mather, W. H., Tsimring, L. S., and Hasty, J. (2008). A fast, robust and tunable synthetic gene oscillator. *Nature* 456, 516–519. doi: 10.1038/nature07389
- Sun, J., Garibaldi, J. M., and Hodgman, C. (2012). Parameter estimation using metaheuristics in systems biology: a comprehensive review. *IEEE/ACM Trans. Comput. Biol. Bioinform.* 9, 185–202. doi: 10.1109/TCBB.2011.63
- Tigges, M., Marquez-Lago, T. T., Stelling, J., and Fussenegger, M. (2009). A tunable synthetic mammalian oscillator. *Nature* 457, 309–312. doi: 10.1038/nature07616
- Tolla, D. A., Kiley, P. J., Lomnitz, J. G., and Savageau, M. A. (2015). Design principles of a conditional futile cycle exploited for regulation. *Mol. Biosyst.* 11, 1841–1849. doi: 10.1039/C5MB00055F
- Tolla, D. A., and Savageau, M. A. (2010). Regulation of aerobic-to-anaerobic transitions by the FNR cycle in *Escherichia coli*. *J. Mol. Biol.* 397, 893–905. doi: 10.1016/j.jmb.2010.02.015
- Tolla, D. A., and Savageau, M. A. (2011). Phenotypic repertoire of the FNR regulatory network in *Escherichia coli*. *Mol. Microbiol.* 79, 149–165. doi: 10.1111/j.1365-2958.2010.07437.x
- Vanderbei, R. J. (1996). *Linear Programming: Foundations and Extensions*. New York, NY: Springer.
- van der Walt, S., Colbert, S. C., and Varoquaux, G. (2011). The NumPy array: a structure for efficient numerical computation. *Comput. Sci. Eng.* 13, 22–30. doi: 10.1109/MCSE.2011.37
- Voit, E. O. (2013). Biochemical systems theory: a review. *ISRN Biomath.* 2013:e897658. doi: 10.1155/2013/897658
- Wyman, J. Jr. (1964). “Linked functions and reciprocal effects in hemoglobin: a second look,” in *Advances in Protein Chemistry*, eds C. B. Anfinsen, M. L. A. John, T. Edsall, and F. M. Richards (Academic Press), 223–286. Available online at: <http://www.sciencedirect.com/science/article/pii/S0065323308601904>
- Yang, X. (2002). Generalized form of Hurwitz-Routh criterion and Hopf bifurcation of higher order. *Appl. Math. Lett.* 15, 615–621. doi: 10.1016/S0893-9659(02)80014-3

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2016 Lomnitz and Savageau. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



Evolution of Centrality Measurements for the Detection of Essential Proteins in Biological Networks

Mahdi Jalili¹, Ali Salehzadeh-Yazdi^{1,2*}, Shailendra Gupta^{2,3}, Olaf Wolkenhauer², Marjan Yaghmaie¹, Osbaldo Resendis-Antonio⁴ and Kamran Alimoghaddam^{1*}

¹ Hematology, Oncology and Stem Cell Transplantation Research Center, Tehran University of Medical Sciences, Tehran, Iran,

² Department of Systems Biology and Bioinformatics, University of Rostock, Rostock, Germany, ³ CSIR-Indian Institute of Toxicology Research, Lucknow, India, ⁴ Human Systems Biology Laboratory, RAI-UNAM and INMEGEN, Mexico City, Mexico

Keywords: essentiality, centrality, biological network, topological network analysis, biological centrality

INTRODUCTION

Current breakthroughs in high-throughput technologies have propelled the development of databases that systematically store knowledge of how genes, proteins, and metabolites interact. To elucidate the mechanisms of molecular interaction, such data can be represented through networks where nodes are biological entities (e.g., gene, protein, miRNA, transcription factor, and metabolites) and edges are associations/interactions between them (e.g., co-expression, signaling, regulation, and physical interaction). One approach to use such networks is to analyze their topological structure and try to relate this to biological function.

Topological analysis hints at the possible behavior of a network in the regulation of biological processes or phenotypes and help in unveiling the core mechanisms. Broadly speaking, topological parameters can be used to explore: (1) collective behaviors (global properties such as diameter, small-world and scale-free properties of a network), (2) subnetwork behaviors (functional motif discovery), and (3) individual behaviors (prioritization of important nodes by centrality indices) of various network components (Ma and Gao, 2012).

One of the first attempts found in the literature considered centrality related to lethality, and is known as the centrality-lethality rule proposed by Jeong et al. indicating a positive correlation between connectivity and indispensability in the yeast protein-protein interaction map (Jeong et al., 2001). Similarly, Wagner and Fell analyzed the structure of a large metabolic network of *E. coli* using metabolite node degree and shortest mean path length and observed small world like properties that follow power-law distributions (Wagner and Fell, 2001). In these two comprehensive studies, an old metric system (centrality index) was applied with different strategies, aiming to answer the following question: Do centrality indices predict the essential nodes in the biological networks?

Remarkably, topological analyses carried out in transcriptional regulatory (TR) and metabolic networks have been a valuable guide to identify those biological components, called essential nodes, that play a major role in vital functional activities for some microorganisms (Resendis-Antonio et al., 2005, 2012). The relationship between nodes topological features, such as their degree, and their essentiality remains however debated (Coulomb et al., 2005).

Prediction of essential proteins is a challenging task because it needs experimental approaches that are expensive, time-consuming, and laborious (Zhong et al., 2013; Li et al., 2014). To optimize the search of essential nodes in biological networks, a series of computational methods that include topological criteria have been proposed. In this paper, we review the cutting edge computational methods by categorizing them according to their underlying strategies to identify

OPEN ACCESS

Edited by:

Alberto Marin-Sanguino,
Technische Universität München,
Germany

Reviewed by:

Marc Santolini,
Northeastern University, USA

*Correspondence:

Ali Salehzadeh-Yazdi
ali.salehzadeh-yazdi@uni-rostock.de
Kamran Alimoghaddam
alimgh@tums.ac.ir

Specialty section:

This article was submitted to
Systems Biology,
a section of the journal
Frontiers in Physiology

Received: 26 August 2015

Accepted: 12 August 2016

Published: 26 August 2016

Citation:

Jalili M, Salehzadeh-Yazdi A, Gupta S,
Wolkenhauer O, Yaghmaie M,
Resendis-Antonio O and
Alimoghaddam K (2016) Evolution of
Centrality Measurements for the
Detection of Essential Proteins in
Biological Networks.
Front. Physiol. 7:375.
doi: 10.3389/fphys.2016.00375

essential components. In each case, we discuss their predictive experimental power and identify shortcomings.

FIRST STRATEGY: USE OF INDIVIDUAL CLASSICAL CENTRALITY INDEX

The most commonly used centrality index is the degree centrality which is calculated as the number of direct connections to a node. Many studies suggested that highly connected nodes or “hubs” are more likely to be essential (Hahn and Kern, 2005; Joyce and Palsson, 2008). For instance, in 2005, Hahn and Kern compared centrality and essentiality in yeast, worm and fly PPI networks and concluded that a protein connectivity has an effect on the probability of being essential (Hahn and Kern, 2005). Nevertheless, high connectivity does not necessarily imply its essentiality. In 2005, Mahadevan and Palsson, indicated that in genome-scale metabolic models of *E. coli*, *S. cerevisiae*, and *Geobacter sulfurreducens*, the essentiality is not correlated with node connectivity (Mahadevan and Palsson, 2005). In addition, in 2007, Tew et al. concluded that in the PPI network a low-connectivity node could also be considered as essential (Tew et al., 2007). To improve upon this, other metrics were suggested to predict essential genes. Thus, almost all classic centrality indices (Freeman, 1979) that were developed for characterizing social networks (such as the degree, closeness, and betweenness centralities) were applied to biological networks. For instance, in 2004, Koschützki and Schreiber applied five centrality indices (degree, eccentricity, closeness, random walk betweenness, and Bonacich's eigenvector) to the PPI network of *Homo sapiens* and gene regulatory network of *E. coli*. They showed that eccentricity and eigenvector are highly correlated in the PPI network while within the TR network a strong correlation between eigenvector and closeness was observed (Koschützki and Schreiber, 2004). Betweenness centrality is based upon the frequency with which a node lies between the shortest communication path of all other possible pairs of nodes within a network and highlights the gatekeepers of communication within the network. Eccentricity centrality of a node is calculated as the reciprocal of the maximum of shortest path lengths from that node to all other nodes in the network. Thus, the node with highest eccentric centrality is considered as the most central node in a network. In contrast the closeness centrality is measured by the reciprocal of sum of the geodesic distances from that node to all other nodes in the network. The basic idea behind the eigenvector centrality of a node was the assumption that centrality index of a node is not only determined by its position in the network but also by the neighboring nodes. Overall degree, betweenness and closeness centrality measurements were among the most common topological parameters investigated in terms of biological network analyses. Potapov et al. introduced a new centrality measurement, named pairwise disconnectivity index, to qualify the importance of individual nodes and/or interactions for sustaining the communications between connected pairs of nodes in a directed network (Potapov et al., 2008). The authors discussed some of the limitations of the betweenness centrality index, mainly the identification of the shortest path

for the communication between a pair of nodes. They argued that the importance of a path does not depend on the length but on other factors, such as the concentration of the species, rate constant etc. Thus, even the longer path can be faster and efficient in biological scenarios. Moreover, the peripheral nodes were not considered. However, in 2014, Raman et al. analyzed the PPI network of a diverse set of 20 organisms. They computed parameters such as degree, betweenness, closeness, and pairwise disconnectivity indices and demonstrated that degree and betweenness centralities correlate with lethality in many organisms but closeness and pairwise disconnectivity indices are not strong indicators of essentiality (Raman et al., 2014).

SECOND STRATEGY: COMBINATION OF CLASSICAL CENTRALITY MEASURES

Some researchers have also attempted to combine the individual centrality matrices to achieve more accurate results. They believe that a single measure of centrality does not solely predict the essential nodes in biological networks. Therefore, combining different centrality indices could yield better results. Examples of such studies include the work of Gabriel del Rio et al. in 2009 on the prediction of essential genes using a new score based on the combination of two or more existing centrality indices (del Rio et al., 2009). They analyzed 16 different centrality measures on 18 reconstructed metabolic networks for *S. cerevisiae* and explained that no single centrality measure identifies essential genes while the combination of at least two centrality measures achieves a reliable prediction. More specifically, they observed that combining “1/clustering coefficient” with either closeness, eccentricity, 1/eccentricity or radiality resulted in significant prediction of essential genes while no improvement was achieved when three or four centrality measures were combined together (del Rio et al., 2009). Wang et al. performed principal component analysis (PCA) to combine eight centralities, and generated a new integrative node importance measure, structurally dominant proteins index, to find more important nodes in the PPI networks. The proposed integrative measure is strongly correlated with eigenvector, semi-local, network motif, degree, and betweenness measures (Wang et al., 2014). The most recent study, named composite centrality, offered a unified scale to measure node, and edge centralities for general weighted and direct complex evolving networks (Joseph and Chen, 2014).

THIRD STRATEGY: USE OF NOVEL CENTRALITY CONCEPTS

In addition to the use of individual classical centrality measures and their combinations to identify essential/lethal nodes in biological networks, new indices were designed using other features associated with nodes in biological networks. For instance, Yu et al. in 2004 introduced the notion of marginal essentiality which states that the essentiality of a gene is directly associated to its connectivity and the number of functions of

that gene (Yu et al., 2004). Estrada and Rodriguez-Velazquez, in 2005 proposed a new index, subgraph centrality (SC) which characterizes the contribution of each node in all subgraphs of a network. The authors claimed that SC index is better in discriminating the nodes of a network than alternate classical measures such as degree, closeness, betweenness, and eigenvector centralities and is more highly correlated with the lethality of individual proteins removed from the proteome (Estrada and Rodriguez-Velazquez, 2005). Tew et al. defined a functional centrality as the topological centrality within a subnetwork of proteins with similar functions, called neighborhood functional centrality (NFC). NFC predicted the lethal proteins in four *S. cerevisiae* PPI datasets and was able to detect low connectivity lethal proteins that were previously undetected by conventional methods (Tew et al., 2007). Then, Koschützki and Schreiber demonstrated that motif-based centralities yield better results in gene regulatory networks (Koschützki and Schreiber, 2008). Efforts were made to better predict and improve the existing methods for new insights of centrality usage in biology. For example, Hart et al. used an unsupervised probabilistic scoring scheme on large-scale yeast mass-spectrometry data, emphasizing that essentiality is the product of protein complexes rather than individual proteins (Hart et al., 2007). Piraveenan et al. used topological connectivity, as well as the percolation states of individual nodes in network percolation scenarios (such as infection transmission in a social network of individuals) to quantify relative impact of nodes (Piraveenan et al., 2013). Simko and Csermely applied game centrality to design more competent interventions in cellular networks (Simko and Csermely, 2013), and Szalay and Csermely developed perturbation centrality to provide a large variety of novel options to assess signaling, drug action, environmental, and social interventions (Szalay and Csermely, 2013). Wuchty recently determined minimum dominating sets (MDSet) as optimized subsets of proteins that play a role in the control of the underlying networks by enabling remaining proteins to be reached in one step. MDSet are enriched with essential, cancer-related, and virus-targeted genes. The author also compared the MDSet proteins with hub proteins and showed a higher impact of MDSet proteins on network resilience (Wuchty, 2014).

FOURTH STRATEGY: INTEGRATION OF OMICS DATA WITH CENTRALITY MEASURES

Until now, we reviewed how mathematical combinations of various centralities generated from complex networks can predict essential genes (Roy, 2012). It seems that the integration of biological knowledge into topological features could create an improved centrality index to find essential nodes. Some studies have also been done in that direction; in 2010, Li et al. improved the prediction of essential proteins 20% more than closeness and subgraph centralities by construction of a weighted PPI based on the combination of logistic regression-based model and function similarity (Li et al., 2010). Li et al. in 2012 introduced and validated a new centrality measure (PeC) by integration of gene expression into the yeast PPI network. In this new

method, a weighting of the PPI network was proposed based on the probability of two proteins to be co-clustered and co-expressed in a given biological scenario. PeC predicted the essential proteins significantly better than the other previously proposed 15 centrality measures: degree, betweenness, closeness, subgraph, eigenvector, information, bottle neck, density of maximum neighborhood component, local average connectivity-based method, sum of edge clustering coefficient, range-limited, L-index, leader rank, modularity, and normalized α -centralities. Above all, the enhancement of PeC over the classic centralities (betweenness, closeness, subgraph, eigenvector, and bottle neck centralities) is more than 50% for the first 500 predictions (Li et al., 2012).

Very recently, Jiang et al. in 2015 developed a network-based method named NEST (Network Essentiality Scoring Tool) that improved the performance of centrality over previous related methods. NEST predicted the essential genes according to the expression level of neighbor genes connected in protein interaction network. The results obtained by the current integration showed that the predictive power of essential protein according to this strategy is much better than the classic centralities (Jiang et al., 2015).

DISCUSSION

Essential genes (and their products, proteins) imply an intricate role in a cell survival and development. Topological network analyses provide opportunities for essential nodes prediction, evaluation of disease genes, and the discovery of potential drug targets (Rosamond and Allsop, 2000). Inspired by previous works in social network analysis (Freeman, 1979; Borgatti et al., 2009), it was assumed that centrality measures could predict essential nodes and several strategies have been offered to find out the relative importance of a node in complex biological networks. However, the structure of biological networks differs fundamentally from social networks especially with respect to modularity (Newman and Park, 2003). Another issue is the dynamic nature of biological entity relationships. For instance, not all relationships may exist simultaneously even in a perfectly mapped network (Han et al., 2004). Therefore, the results of centrality indices in the prediction of essential nodes were not satisfactory in various studies. One of the proposed solutions is to apply functional methods in this context according to the type of biological networks to be analyzed. Such methods integrating other aspects of biological knowledge could be very helpful. In addition, ranking genes or proteins through more biologically driven features such as physicochemical properties of bio macromolecules, intrinsic disorder property of proteins, co-expression of biological entities, gene clusters, protein complexes, protein localization, gene ontology, enrichment analysis, two-dimensional annotation of genomes, different types of promoters, and epistatic interaction will be of interest. Now that more biological data is available, it is time to improve over the pure topological measures and redefine the concept of centrality on the basis of specific properties of biological functions. A systematic look into the biological concepts is required; implying that several features could be involved and their combination would result in

an improved biological centrality. More detailed analyses and discussions among researchers are needed to decide upon the parameters to be combined with different centrality measures for the prediction of essential genes in context specific biological networks. There is no particular reason to expect an exact match between network topology and biological functions. As such these tools provide the basis for “intelligent guessing.” In view of the complexity of biological

networks and the difficulties to generate experimental data for other analyses, providing hints can prove already very useful.

AUTHOR CONTRIBUTIONS

All authors listed, have made substantial, direct and intellectual contribution to the work, and approved it for publication.

REFERENCES

- Borgatti, S. P., Mehra, A., Brass, D. J., and Labianca, G. (2009). Network analysis in the social sciences. *Science* 323, 892–895. doi: 10.1126/science.1165821
- Coulomb, S., Bauer, M., Bernard, D., and Marsolier-Kergoat, M.-C. (2005). Gene essentiality and the topology of protein interaction networks. *Proc. Biol. Sci.* 272, 1721–1725. doi: 10.1098/rspb.2005.3128
- del Rio, G., Koschützki, D., and Coello, G. (2009). How to identify essential genes from molecular networks? *BMC Syst. Biol.* 3:102. doi: 10.1186/1752-0509-3-102
- Estrada, E., and Rodriguez-Velazquez, J. A. (2005). Subgraph centrality in complex networks. *Phys. Rev. E* 71:056103. doi: 10.1103/PhysRevE.71.056103
- Freeman, L. C. (1979). Centrality in social networks conceptual clarification. *Soc. Networks* 1, 215–239. doi: 10.1016/0378-8733(78)90021-7
- Hahn, M. W., and Kern, A. D. (2005). Comparative genomics of centrality and essentiality in three eukaryotic protein-interaction networks. *Mol. Biol. Evol.* 22, 803–806. doi: 10.1093/molbev/msi072
- Han, J.-D. J., Bertin, N., Hao, T., Goldberg, D. S., Berriz, G. F., Zhang, L. V., et al. (2004). Evidence for dynamically organized modularity in the yeast protein-protein interaction network. *Nature* 430, 88–93. doi: 10.1038/nature02555
- Hart, G. T., Lee, I., and Marcotte, E. R. (2007). A high-accuracy consensus map of yeast protein complexes reveals modular nature of gene essentiality. *BMC Bioinformatics* 8:236. doi: 10.1186/1471-2105-8-236
- Jeong, H., Mason, S. P., Barabasi, A. L., and Oltvai, Z. N. (2001). Lethality and centrality in protein networks. *Nature* 411, 41–42. doi: 10.1038/35075138
- Jiang, P., Wang, H., Li, W., Zang, C., Li, B., Wong, Y. J., et al. (2015). Network analysis of gene essentiality in functional genomics experiments. *Genome Biol.* 16, 1–10. doi: 10.1186/s13059-015-0808-9
- Joseph, A. C., and Chen, G. (2014). Composite centrality: a natural scale for complex evolving networks. *Physica D Nonlinear Phenomena* 267, 58–67. doi: 10.1016/j.physd.2013.08.005
- Joyce, A. R., and Palsson, B. Ø. (2008). Predicting gene essentiality using genome-scale *in silico* models. *Methods Mol. Biol.* 416, 433–457. doi: 10.1007/978-1-59745-321-9_30
- Koschützki, D., and Schreiber, F. (2008). Centrality analysis methods for biological networks and their application to gene regulatory networks. *Gene Regul. Syst. Bio.* 2, 193–201.
- Koschützki, D., and Schreiber, F. (2004). “Comparison of centralities for biological networks,” in *German Conference on Bioinformatics* (Bielefeld).
- Li, M., Wang, J., Wang, H., and Pan, Y. (2010). “Essential proteins discovery from weighted protein interaction networks,” in *Bioinformatics Research and Applications*, eds M. Borodovsky, J. P. Gogarten, T. M. Przytycka, and S. Rajasekaran (Storrs, CT: Springer), 89–100.
- Li, M., Zhang, H., Wang, J. X., and Pan, Y. (2012). A new essential protein discovery method based on the integration of protein-protein interaction and gene expression data. *BMC Syst. Biol.* 6:15. doi: 10.1186/1752-0509-6-15
- Li, M., Zheng, R., Zhang, H., Wang, J., and Pan, Y. (2014). Effective identification of essential proteins based on priori knowledge, network topology and gene expressions. *Methods* 67, 325–333. doi: 10.1016/j.ymeth.2014.02.016
- Ma, X., and Gao, L. (2012). Biological network analysis: insights into structure and functions. *Brief. Funct. Genomics* 11, 434–442. doi: 10.1093/bfpg/els045
- Mahadevan, R., and Palsson, B. (2005). Properties of metabolic networks: structure versus function. *Biophys. J.* 88, L07–L09. doi: 10.1529/biophysj.104.055723
- Newman, M. E., and Park, J. (2003). Why social networks are different from other types of networks. *Phys. Rev. E* 68:036122. doi: 10.1103/PhysRevE.68.036122
- Piraveenan, M., Prokopenko, M., and Hossain, L. (2013). Percolation centrality: quantifying graph-theoretic impact of nodes during percolation in networks. *PLoS ONE* 8:e53095. doi: 10.1371/journal.pone.0053095
- Potapov, A. P., Goemann, B., and Wingender, E. (2008). The pairwise disconnectivity index as a new metric for the topological analysis of regulatory networks. *BMC Bioinformatics* 9:227. doi: 10.1186/1471-2105-9-227
- Raman, K., Damaraju, N., and Joshi, G. K. (2014). The organisational structure of protein networks: revisiting the centrality-lethality hypothesis. *Syst. Synth. Biol.* 8, 73–81. doi: 10.1007/s11693-013-9123-5
- Resendis-Antonio, O., Freyre-González, J. A., Menchaca-Méndez, R., Gutiérrez-Ríos, R. M., Martínez-Antonio, A., Ávila-Sánchez, C., et al. (2005). Modular analysis of the transcriptional regulatory network of *E. coli*. *Trends Genetics* 21, 16–20. doi: 10.1016/j.tig.2004.11.010
- Resendis-Antonio, O., Hernández, M., Mora, Y., and Encarnación, S. (2012). Functional modules, structural topology, and optimal activity in metabolic networks. *PLoS Comput. Biol.* 8:e1002720. doi: 10.1371/journal.pcbi.1002720
- Rosamond, J., and Allsop, A. (2000). Harnessing the power of the genome in the search for new antibiotics. *Science* 287, 1973–1976. doi: 10.1126/science.287.5460.1973
- Roy, S. (2012). Systems biology beyond degree, hubs and scale-free networks: the case for multiple metrics in complex networks. *Syst. Synth. Biol.* 6, 31–34. doi: 10.1007/s11693-012-9094-y
- Simko, G. I., and Csérmely, P. (2013). Nodes having a major influence to break cooperation define a novel centrality measure: game centrality. *PLoS ONE* 8:e67159. doi: 10.1371/journal.pone.0067159
- Szalay, K. Z., and Csérmely, P. (2013). Perturbation centrality and Turbine: a novel centrality measure obtained using a versatile network dynamics tool. *PLoS ONE* 8:e78059. doi: 10.1371/journal.pone.0078059
- Tew, K. L., Li, X. L., and Tan, S. H. (2007). Functional centrality: detecting lethality of proteins in protein interaction networks. *Genome Inform.* 19, 166–177. doi: 10.1142/9781860949852_0015
- Wagner, A., and Fell, D. A. (2001). The small world inside large metabolic networks. *Proc. Biol. Sci.* 268, 1803–1810. doi: 10.1098/rspb.2001.1711
- Wang, P., Yu, X., and Lu, J. (2014). Identification and evolution of structurally dominant nodes in protein-protein interaction networks. *IEEE Trans. Biomed. Circ. Syst.* 8, 87–97. doi: 10.1109/TBCAS.2014.2303160
- Wuchty, S. (2014). Controllability in protein interaction networks. *Proc. Natl. Acad. Sci. U.S.A.* 111, 7156–7160. doi: 10.1073/pnas.1311231111
- Yu, H., Greenbaum, D., Lu, H. X., Zhu, X., and Gerstein, M. (2004). Genomic analysis of essentiality within protein networks. *Trends genetics* 20, 227–231. doi: 10.1016/j.tig.2004.04.008
- Zhong, J., Wang, J., Peng, W., Zhang, Z., and Pan, Y. (2013). Prediction of essential proteins based on gene expression programming. *BMC Genomics* 14 Suppl 4:S7. doi: 10.1186/1471-2164-14-S4-S7

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2016 Jalili, Salehzadeh-Yazdi, Gupta, Wolkenhauer, Yaghmaie, Resendis-Antonio and Alimoghaddam. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



Logical Modeling and Dynamical Analysis of Cellular Networks

Wassim Abou-Jaoudé¹, Pauline Traynard¹, Pedro T. Monteiro^{2,3}, Julio Saez-Rodriguez⁴, Tomáš Helikar⁵, Denis Thieffry¹ and Claudine Chaouiya^{3*}

¹ Computational Systems Biology Team, Institut de Biologie de l'Ecole Normale Supérieure, CNRS UMR8197, INSERM U1024, Ecole Normale Supérieure, PSL Research University, Paris, France, ² INESC-ID/Instituto Superior Técnico, University of Lisbon, Lisbon, Portugal, ³ Instituto Gulbenkian de Ciência, Oeiras, Portugal, ⁴ Faculty of Medicine, Joint Research Centre for Computational Biomedicine, RWTH Aachen University, Aachen, Germany, ⁵ Department of Biochemistry, University of Nebraska-Lincoln, Lincoln, NE, USA

The logical (or logic) formalism is increasingly used to model regulatory and signaling networks. Complementing these applications, several groups contributed various methods and tools to support the definition and analysis of logical models. After an introduction to the logical modeling framework and to several of its variants, we review here a number of recent methodological advances to ease the analysis of large and intricate networks. In particular, we survey approaches to determine model attractors and their reachability properties, to assess the dynamical impact of variations of external signals, and to consistently reduce large models. To illustrate these developments, we further consider several published logical models for two important biological processes, namely the differentiation of T helper cells and the control of mammalian cell cycle.

Keywords: regulatory and signaling networks, logical modeling, discrete dynamics, attractors, reachability analysis, simulation, T cells activation and differentiation, cell cycle control

OPEN ACCESS

Edited by:

Rui Alves,
Universitat de Lleida, Spain

Reviewed by:

Monika Heiner,
Brandenburg Technical University
Cottbus-Senftenberg, Germany
Noriko Hiroi,
Keio University, Japan

*Correspondence:

Claudine Chaouiya
chaouiya@igc.gulbenkian.pt

Specialty section:

This article was submitted to
Systems Biology,
a section of the journal
Frontiers in Genetics

Received: 30 January 2016

Accepted: 12 May 2016

Published: 31 May 2016

Citation:

Abou-Jaoudé W, Traynard P,
Monteiro PT, Saez-Rodriguez J,
Helikar T, Thieffry D and Chaouiya C
(2016) Logical Modeling and
Dynamical Analysis of Cellular
Networks. *Front. Genet.* 7:94.
doi: 10.3389/fgene.2016.00094

1. INTRODUCTION

As computational modeling is increasingly recognized as a necessary and valuable approach to understand dynamical features of complex biological processes, the logical framework has proved to be particularly successful to model and analyze regulatory and signaling networks (Samaga and Klamt, 2013; Albert and Thakar, 2014; Le Novère, 2015; Naldi et al., 2015). Back in 1961, following the discovery of specific gene regulation mechanisms and the delineation of the first regulatory circuits in bacteria (Jacob and Monod, 1961; Monod and Jacob, 1961), several researchers proposed to use Boolean algebra to model cellular circuits. Mitoyosi Sugita was the first to present an explicit modeling of bacterial genetic circuits with symbolic logic, applying the methods and tools of mathematics and electronics, and coining the term molecular automaton (Sugita, 1963). Soon after, Stuart Kauffman engaged in a thorough analysis of the dynamical properties of generic Boolean network models, using a synchronous update and focusing on asymptotical properties (Kauffman, 1969; Glass and Kauffman, 1973). In parallel, René Thomas rather addressed the modeling of specific regulatory circuits, in particular the network controlling lysis-lysogeny decision in bacteriophage lambda, using an asynchronous update, and progressively refining the logical formalism with the introduction of multi-valued variables, the explicit consideration of threshold values, the definition of logical parameters, etc. (Thomas, 1973; Thomas et al., 1976; Thomas, 1978). By and large, the studies of Kauffman and Thomas converged in showing that alternative stable states (or more generally alternative attractors) can be associated with different cell types, and that logical state transitions can be associated with gene expression changes over

time. While Kauffman emphasized how connectivity and specific kinds of logical functions have an impact on the asymptotic network behavior, Thomas focused more specifically on the dynamical roles of simple, positive vs. negative regulatory circuits embedded in more complex networks. Altogether, these contributions laid the foundation for a wealth of studies demonstrating the versatility and power of logical modeling in molecular biology and beyond (see e.g., Thomas and D'Ari, 1990; Kauffman, 1993).

Briefly, in a logical model, each component is associated with a discrete variable, which is a logical (often Boolean i.e., binary) abstraction of its level of activity (or concentration). A logical function defines the next value of this variable, depending on the current levels of the regulators of that component. Such a model defines a discrete dynamical system where the state of the network (the component levels) evolves stepwise. Besides scalability (logical models with few hundreds components have been simulated), the appeal of this framework relies on its qualitative nature, as kinetic parameters and other precise knowledge about the molecular mechanisms at stake are not required. Despite this coarse grained abstraction, the resulting behaviors presumably capture the most salient properties of the modeled systems (Samaga and Klamt, 2013; Albert and Thakar, 2014; Le Novère, 2015). As a matter of fact, the logical framework proved useful in a wide range of biological applications: cell differentiation in developmental processes (for instance, drosophila development as in González et al., 2008; Sánchez et al., 2008; Fauré et al., 2014), haematopoiesis (Bonzanni et al., 2013), T lymphocyte activation and differentiation (see Section 5.1), cell cycle control (see Section 5.2) and more generally cell fate decisions such as proliferation, growth arrest, apoptosis, senescence, etc. (see e.g., Schlatter et al., 2009; Grieco et al., 2013; Mombach et al., 2014; Cohen et al., 2015).

Alternative modeling frameworks explicitly refer to sets of reaction rules (denoting molecular consumption and production processes) to model and analyze cellular networks (see Le Novère, 2015 for further details and references). In this respect, a logical model can be considered as an abstraction focusing on signed interactions denoting positive or negative influences between network components (defining the regulatory graph, which is completed by logical rules specifying the compositional effects of these influences). The logical framework is thus primarily used for signaling and gene regulation modeling.

For a general overview of the logical modeling of biological networks, we refer to existing reviews (Samaga and Klamt, 2013; Albert and Thakar, 2014; Le Novère, 2015; Naldi et al., 2015). Here, we emphasize the versatility of the logical formalism, as well as the relevance of a range of methods and tools. We first present formal definitions of (multi-valued) models and their associated dynamics, depending on a variety of updating schemes. As attractors and their reachability are of utmost interest when analyzing models of biological networks (see e.g., Huang et al., 2009), we particularly focus on approaches to determine model attractors and their reachability properties, as well as on the impact of variations of external signals on model behaviors. To demonstrate the relevance of the logical modeling and of the associated methodological advances, we survey several

published logical models dealing with two important biological processes: (i) the activation and differentiation of T cells, and (ii) the control of cell proliferation.

The regulatory network controlling mammalian T helper (Th) lymphocyte activation and differentiation is of particular interest from the modeling point of view. First, this system has been largely studied experimentally, leading to the identification of many of the key molecular components involved. Furthermore, Th cell activation and differentiation are controlled by complex and intertwined intracellular signaling pathways and regulatory circuits, which ultimately enable the differentiation of Th cells into multiple functional subtypes, depending on the signals present in their microenvironment.

Also challenging and well studied are the networks controlling the initiation of cell division and the progression of cells along the main phases of mitotic cell cycles. Initially investigated in model organisms such as budding and fission yeasts, these networks have been deciphered in various other species, up to mammals. Models have been built to assess the implementation of the various cell cycle check points and the achievement of coordinated and robust oscillations in the activities of molecular components. Moreover, as defects of the cell cycle engine are one of the bases of cancer, many studies currently focus on mammalian cell cycle control networks.

In Section 2, we formally introduce the basics of the logical formalism and its main variants. The core of this paper demonstrates the assets of the framework with advanced methods and tools to analyze behavioral properties (Section 3), as well as to support data integration into models (Section 4). Finally, we illustrate these assets on T cell signaling and cell cycle control networks (Section 5).

2. FUNDAMENTALS OF THE LOGICAL FORMALISM

We formally introduce the logical framework, defining models and their dynamics. Most common variants are presented, in particular regarding updating schemes and their impacts on dynamical properties. A selection of computational tools is then briefly presented.

2.1. Model Definition

The basic concepts presented in this section are illustrated in **Figure 1**. A logical model $(\mathcal{G}, \mathcal{K})$ of a regulatory network is defined by:

- A set of n regulatory components $\mathcal{G} = \{g_1, g_2, \dots, g_n\}$, each g_i being associated with an integer variable, which takes its values in $\{0, \dots, \max_i\}$, defining a discrete mapping of the range of the component functional levels (of activity or concentration). The (finite) state space \mathcal{S} is defined as the cartesian product $\prod_{i=1, \dots, n} \{0, \dots, \max_i\}$ and a model state is thus a vector $g = (g_1, \dots, g_n)$.
- For each g_i , a discrete function \mathcal{K}_i defines its values, depending on the model states: $\mathcal{K}_i : \mathcal{S} \rightarrow \{0, \dots, \max_i\}$. The transition function $\mathcal{K} : \mathcal{S} \rightarrow \mathcal{S}$ with $\mathcal{K}(g) = (\mathcal{K}_1(g), \dots, \mathcal{K}_n(g))$ thus

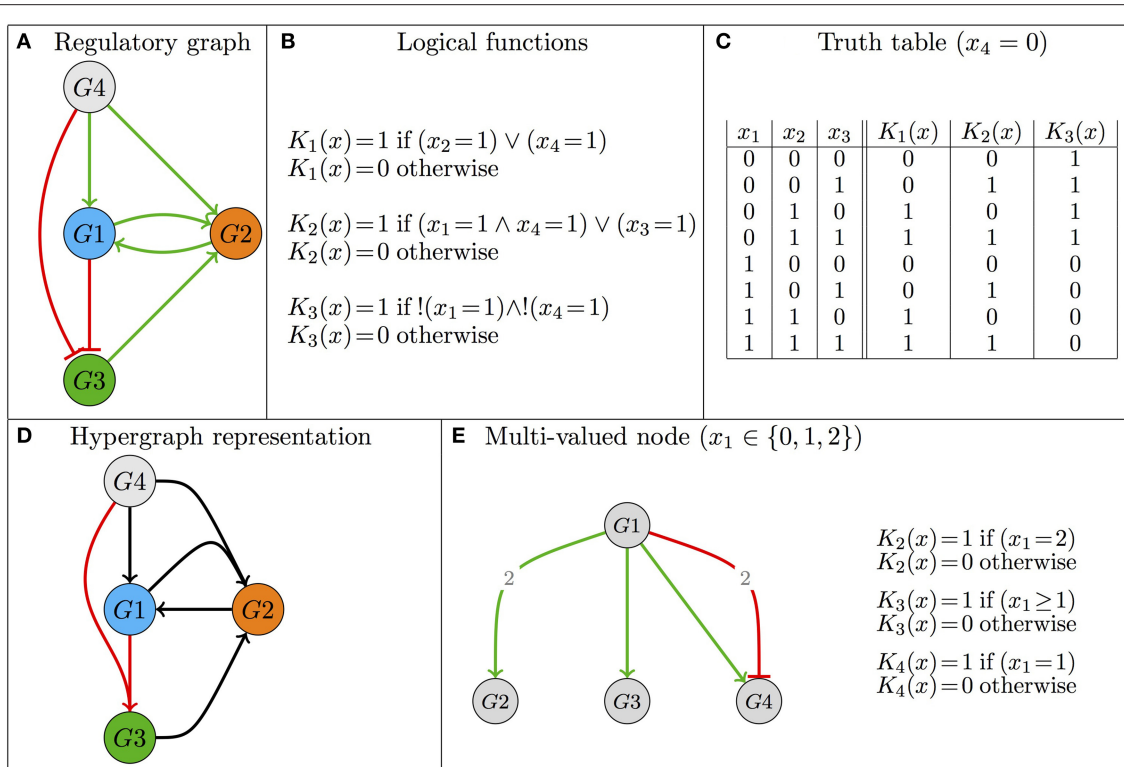


FIGURE 1 | Illustration of the basics of the logical formalism—Model definition. (A) The regulatory graph defines the topology of the regulatory structure, where nodes denote regulatory components and edges represent regulatory effects (activations are denoted by green edges, whereas inhibitions are represented in red). (B,C) The evolution of the variables associated with the regulatory components is defined by the logical functions, which are written in the form of logical formulas or, alternatively, in the form of truth tables. “ \wedge ,” “ \vee ,” and “ \neg ” stand for the logical operators AND, OR and NOT, respectively. Note that the regulatory graph in (A) can be recovered from the logical functions defined in (B,C), but the reverse is not true (see main text). (D) Hypergraph as an alternative definition of the Boolean model of (A–C) (merged arrows denote AND operator). (E) Example motivating the introduction of a multi-valued variable; here G1 activates G2 and G3 at different thresholds and activates G4 when it is at level 1, but inhibits it at level 2 (see also **Supplementary Figure S1**).

defines the model behavior, but also the underlying regulatory graph (see below).

While a Boolean discretization is generally enough (i.e., $\max_i = 1$ for all i), a regulatory component may operate at different levels on distinct targets, or yet, depending on its level, may have different effects on a given target. In such cases, it is necessary to consider a multi-valued variable whose maximal value is greater than 1 (see **Figure 1E**). Note that the discrete functions K_i are referred to as *logical functions*, even in the case of multi-valued variables. This denomination originates in Thomas and Snoussi’s seminal work defining their *generalized kinetic logic* (Thomas and D’Ari, 1990).

The regulatory graph, denoted $(\mathcal{G}, \mathcal{R})$, is often available early on. It encompasses nodes denoting model components (regulatory components, elements of \mathcal{G}), along with signed, directed edges, denoting regulatory activations or inhibitions (elements of \mathcal{R}). The logical rules precisely encode these interactions. In other words, $(\mathcal{G}, \mathcal{R})$ can be deduced from \mathcal{K} . Note, however, that several sets of logical rules can be compliant with a regulatory graph, which therefore defines a family of logical models.

There is a *functional* interaction from g_j to g_i (denoted $(g_j, g_i) \in \mathcal{R}$) if and only if there exists a pair of neighboring states that only differ on the value of g_j and for which the function K_i takes a different value, thus indicating that a variation of g_j has an effect on the value of its target g_i . More formally, assuming for simplicity that g_j is a Boolean variable, $(g_j, g_i) \in \mathcal{R}$ if and only if there exist two states $g = (g_1, \dots, g_{j-1}, 1, g_{j+1}, \dots, g_n)$ and $\bar{g} = (g_1, \dots, g_{j-1}, 0, g_{j+1}, \dots, g_n)$ such that $K_i(g) \neq K_i(\bar{g})$. Moreover, if $K_i(\bar{g}) < K_i(g)$, this interaction is an activation (because when $g_j = 0$ as in state \bar{g} , the function K_i defines a lower value for g_i than when $g_j = 1$ as in state g), otherwise it is an inhibition.

Specific classes of Boolean regulatory functions have been considered in the literature. The simplest specifies that a component is activated (its associated variable tends to 1) in the presence of at least one of its activators and in the absence of all of its inhibitors (e.g., Mendoza and Xenarios, 2006). Threshold networks constitute another popular class of Boolean models, in which the regulatory function is defined by comparing the (possibly weighted) sum of positive and negative regulatory contributions with a specific threshold (Li et al., 2004; Bornholdt, 2008). Finally, relying on the fact that any Boolean function can be written in a disjunctive normal form (a disjunction

of conjunctive clauses, thus using exclusively the operators AND, OR and NOT), an alternative, refined representation uses hypergraphs (Klamt et al., 2009; Samaga and Klamt, 2013, Figure 1D).

2.2. Model Dynamics

A logical model defines a discrete dynamics over its state space S . Given a state g , the transition function \mathcal{K} specifies the possible changes of the model variables: if $\mathcal{K}(g) \neq g$, there is at least one variable g_i called to update toward the target value $\mathcal{K}_i(g)$. Note that multi-valued variables are modified stepwise, i.e., if $\mathcal{K}_i(g)$ differs from the current value of g_i by a value greater than 1, the next value of g_i is increased (or decreased) by 1. If $\mathcal{K}(g) = g$ then g is a stable state, in which each component value is maintained constant. Input components, which typically embody external signals, have no regulators and hence no associated logical rules. They are generally considered as being constant (their values representing a fixed environmental condition). However, how the model evolves upon input variations is of particular interest and is discussed in Section 3.2.

Model dynamics are conveniently represented in terms of *State Transition Graphs* (STG), where nodes denote states, while directed edges represent state transitions (Figure 2). Since the number of states is finite, model simulations always end up in a stable state or in a (potentially branched) cyclic trajectory. Stable states (devoid of transitions to other states) often represent cell differentiated states (cf. Section 5.1) or other kind of relevant, perduring situations. In contrast, cyclic trajectories may denote a biologically relevant periodic behavior, as in the case of cell cycle (cf. Section 5.2) or circadian rhythms. The mathematical counterparts of such asymptotic behaviors are called *attractors*, which are defined in the context of the logical formalism as terminal *Strongly Connected Components* (SCC) of the STG, i.e., maximal sets of mutually reachable states, with no transitions leaving the set. The set of states from which trajectories (exclusively) lead to an attractor is called its (strict) *basin of attraction*. Basins of attraction are particularly relevant since they define the reachable attractor(s) depending on the chosen initial state(s).

Dynamical properties of interest predominantly relate to the existence and reachability of the attractors. These are properties hard to assess in large models because the size of the state space (and thus of the STG) grows exponentially with the number of regulatory components. Section 3 presents several recent methods to identify attractors and to check their reachability properties.

If at state g , several variables are called to change their values (because their current values differ from the values returned by the corresponding logical functions), one has to specify how these changes should be performed. The two most common schemes are the *synchronous* and *asynchronous* updates. According to the first, all the variable updates are performed synchronously (i.e., simultaneously). Hence, the resulting deterministic dynamics defines, at each time step t (or iteration), the successor state of $g(t)$:

$$g(t+1) = \left(g_i(t) + \text{sign}(\mathcal{K}_i(g(t)) - g_i(t)) \right)_{i=1, \dots, n}, \quad (1)$$

where $\text{sign}(p)$ equals to 1 if $p > 0$, -1 if $p < 0$, and 0 otherwise. According to Equation (1), a successor state is defined by increasing or decreasing by 1 all the variables whose current values differ from the values specified by their logical functions. Note that if all the variables are Boolean, this equation can be written simply as $g(t+1) = \mathcal{K}(g(t))$. Given a state g , the synchronous update yields exactly one transition toward a successor state, which can be g itself, if all the variables are stable in g , i.e., $\mathcal{K}_i(g) = g_i$ for all $g_i \in \mathcal{G}$.

In contrast, with the asynchronous update, each variable is modified independently, yielding as many transitions (and successor states) as the number of updated variables (and hence potentially non-deterministic dynamics). At state $g(t)$, for all $g_i \in \mathcal{G}$ such that $\mathcal{K}_i(g(t)) \neq g_i(t)$, an asynchronous successor $g(t+1)$ of $g(t)$ is defined as follows:

$$\begin{aligned} g_i(t+1) &= g_i(t) + \text{sign}(\mathcal{K}_i(g(t)) - g_i(t)), \\ g_j(t+1) &= g_j(t) \text{ for all } j \neq i. \end{aligned} \quad (2)$$

Note that, according to this definition, a stable state has no successor. However, for any updating scheme, one may alternatively consider that a stable state is its own successor (with a self-loop transition).

In the context of asynchronous dynamics, priority classes, deterministic and stochastic schemes have been proposed, taking into account additional knowledge to penalize or discard unrealistic trajectories. Indeed, update classes can be defined, grounded on the nature of the processes involved, e.g., different time scales associated with transcriptional and post-translational processes (Chaves et al., 2005). At each time step (or iteration), the selection of updated variables is directed by their associated priority classes (Fauré et al., 2006), absolute ranks or probabilities (e.g., Albert and Thakar, 2014 and references therein). Generalizing the logical framework with a probabilistic interpretation, a finite Markov chain can be derived from the dynamics of a logical model. Considering the asynchronous update, Stoll et al. (2012) defined continuous or discrete time Markov processes by associating stochastic rates with the updates of the model components, and relied on a Gillespie algorithm to simulate the time evolution of component levels. This allows to get a more quantitative view of the model behavior (cf. Section 5.2). In Cell Collective, synchronous simulations also result in a Markov chain when the input components are associated with a probability (see Section 3.2 and Todd and Helikar, 2012).

When following a unique trajectory (defined by a synchronous update or selecting specific transitions among multiple asynchronous, concurrent trajectories), a natural alternative to the STG consists in displaying the evolution of the individual variables over time (see Figure 2C). To provide a different view of the model behavior, it has been also proposed to consider the mean values \tilde{g}_i of a model variable g_i over a sliding window of (user-defined) length w (Helikar and Rogers, 2009):

$$\forall g_i \in \mathcal{G}, \forall t \geq 0, \tilde{g}_i(t) = \frac{\sum_{0 \leq k < \min(w, t)} g_i(t-k)}{\min(w, t)}. \quad (3)$$

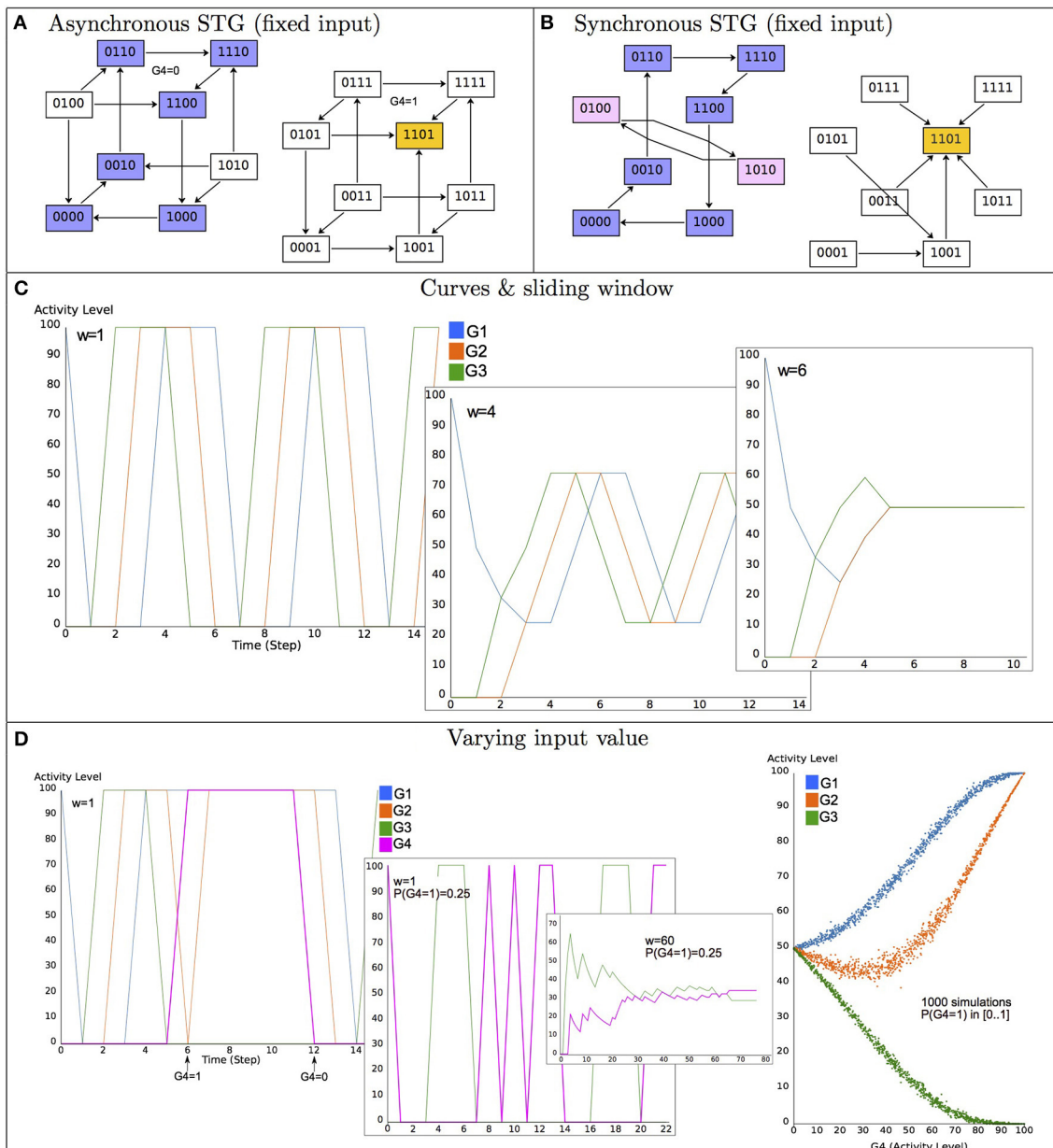


FIGURE 2 | Illustration of the basics of the logical formalism—Model dynamics. (A) The asynchronous State Transition Graph (STG) of the model defined in Figure 1 (A–C), with the input G_4 maintained constant and concurrent transitions from states in which several variables are called to update their values. The yellow state 1101 (i.e., $x_1 = x_2 = x_4 = 1$ and $x_3 = 0$) is a stable state, the set of states in blue corresponds to a cyclic attractor. (B) The synchronous STG in which variables are simultaneously updated; the stable state is conserved, whereas a new terminal cycle appears (in pink). (C) Synchronous dynamics starting from the state 1000 and maintaining the input constant to 0 (activity levels are given in %, from 0 to 100%). For a sliding window of length $w = 1$ (see Equation 3), the curves conform the terminal cycle of (B) (in blue), the four variables oscillate between 0 and 1, with a period of 6; for $w = 4$, the mean values oscillate between 0.25 and 0.75; for $w = 6$, the mean values are constant to 0.5. (D) Illustration of the effect of different input variations (G_4 value). When G_4 is active with a probability 0.25, oscillations of the remaining components are altered (only G_3 values are displayed, for legibility). The plot on the right shows the effect of varying the probability of G_4 activity (from 0 to 1) on the mean values of the remaining components in the long term (i.e., in the attractor).

It is worth recalling that different updating schemes lead to different dynamics, thus impacting related properties (e.g., see Albert and Thakar, 2014). Briefly, compared to the synchronous scheme, asynchronous dynamics are more realistic in accounting for delays between updating orders and their executions. While

stable states are the same for both the synchronous and asynchronous schemes, a striking example of how the resulting dynamics can differ is that of isolated regulatory circuits, for which the synchronous scheme leads to the appearance of additional cyclic attractors (Remy et al., 2003). Not only

cyclical attractors may be different, but reachability properties are also modified. The asynchronous scheme generates concurrent trajectories, some of which are potentially unfeasible in regard to well-grounded choices between concurrent events. Hence, refined asynchronous schemes have been considered, such as priorities, fixed ranks or probabilities, which may also affect attractors and their reachability properties. Indeed, as some trajectories are preempted, transient oscillatory behaviors may be turned into cyclic attractors.

By way of conclusion, beyond the model definition as presented in Section 2.1, modelers need to specify an updating scheme and make this choice explicit when presenting their results. Moreover, model robustness could be assessed by probing different updating schemes and their impacts on attractors and their reachability properties.

2.3. A Selection of Computational Tools

Here, we focus on the software tools used to generate results reported in the remaining sections. The web page <http://colomoto.org/software/> provides a more comprehensive overview of available network modeling tools based on the logical framework.

GINsim (<http://ginsim.org>) supports the definition of multi-valued logical models, under the synchronous, asynchronous and priority updating schemes. Besides the explicit construction of STG (for reasonable sizes, i.e., in the order of a few million states), GINsim provides a number of methods to analyze model properties and supports model exports into various formats, in particular for model checking (see Section 3.1) (Chaouiya et al., 2012; Béranguier et al., 2013).

Cell Collective (<http://cellcollective.org>) is a web-based software with a user friendly interface for model construction, simulation and analyses in a collaborative fashion. Its model repository provides a way for users to directly use and/or expand any of the 50 or so available models. Cell Collective supports Boolean models, considers synchronous updates, stochastic input simulations, and semi-continuous dose-response (input-output) analyses as shown in Section 5.2 (Helikar et al., 2012, 2013b).

CellNetOptimizer (CellNOpt, <http://www.cellnopt.org>) permits to define models of signaling networks as Boolean synchronous models. It further supports constrained fuzzy logic (Morris et al., 2011) and systems of differential equations (Wittmann et al., 2009). CellNOpt specificity is that, starting from a Prior Knowledge Network (i.e., a candidate topology of the signaling network under study), it creates a model by fitting its behavior to high-throughput biochemical data (MacNamara et al., 2012; Terfve et al., 2012).

MaBoSS (<http://maboss.curie.fr>) is a command-line tool simulating continuous/discrete time Markov processes induced by Boolean models (Stoll et al., 2012). Stochastic rates are associated with model component updates and a Gillespie algorithm is used to simulate the time evolution of component levels. Time evolutions of probabilities are estimated and global and semi-global characterizations of the whole system dynamics are further provided.

3. MODEL ANALYSIS

In this section we focus on a selection of methods to assess dynamical properties of logical models. Usage and relevance of these methods are illustrated in Section 5. We refer to Morris et al. (2010), Samaga and Klamt (2013), Albert and Thakar (2014), and Naldi et al. (2015) for further overviews.

3.1. Identifying the Attractors and Analyzing Their Reachability

As previously mentioned, properties of interest relate to attractors and their reachability properties. In small models (up to a dozen components), such properties can be easily recovered directly by constructing and analyzing the State Transition Graph (STG). However, for larger models, a variety of approaches based on different algorithmic techniques and efficient data structures have been proposed to handle the combinatorial explosion of the number of states.

Stable states, which do not depend on updating schemes, are relatively easy to identify because they correspond to the fixed points of the transition function. The algorithm implemented in GINsim relies on (multi-valued) decision diagrams to represent the (Boolean) stability function of each component g_i (which is true iff $\mathcal{K}_i(g) = g_i$). Proper manipulations of this data structure enable the identification of all the stable states of a logical model of up to about hundred components (Naldi et al., 2007).

Identification of complex attractors is harder. Those are composed of several states and depend on the selected updating scheme (cf. **Figure 2**). In a synchronous dynamics, they correspond to terminal, elementary cycles (i.e., closed dynamical cycles in which each state has a unique successor), whose states are fixed points of the p^{th} iterate of \mathcal{K} , for a cycle of length p (note that p is not known in advance). Hence, most existing methods sample or explore the whole STG. Binary Decision Diagrams proved effective to perform such an exploration (Garg et al., 2008). Avoiding exploration of the state space, methods to identify stable subspaces (i.e., regions of the space space in which the model dynamics is trapped and thus contain attractors) have been recently proposed (Zañudo and Albert, 2013; Klarner et al., 2015).

Hierarchical Transition Graphs (HTG) have been defined as STG compactions revealing crucial properties of the dynamics (Béranguier et al., 2013). Briefly, a HTG gathers (i) states that belong to the same SCC, and (ii) states that define trivial SCCs (i.e., if reached once, they cannot be revisited) and from which the same set of attractors and SCCs can be reached (cf. **Supplementary Figure S1**). Hence a HTG provides an informative view of the dynamics in terms of attractors and their basins of attraction.

To quantify attractor reachability, Mendes et al. (2014) presented Avatar, a Monte Carlo simulation algorithm adapted to speed up exit from transient cycles and to identify complex attractors if those are not known beforehand. Avatar allows to estimate the probability of reaching an attractor from an initial state or from any initial state (i.e., sampling the state space) under the assumption of equiprobability of concurrent transitions. In turn, MaBoSS, mentioned in Section 2.3, provides an estimation

of state probabilities over time (cf. Section 5.2), along with further characterizations of the whole dynamics.

Model checking was proposed in the early 1980s to verify a (set of) specification(s) against very large models of hardware and software systems. Since then, methodologies have been improved as well as their ranges of applicability. Notably, in the mid 2000s, model checking started to be applied in Systems Biology, mainly to verify qualitative systems dynamics (e.g., Chabrier and Fages, 2003; Batt et al., 2005; Arellano et al., 2011; Abou-Jaoudé et al., 2015), but also for hybrid systems considering continuous time or continuous state variables (e.g., Hinton et al., 2006; Clarke et al., 2008; see also Brim et al., 2013 for an overview). A model checker verifies whether a model of a system satisfies a set of properties, answering true/false for each property. The dynamics is represented as a specific transition system and properties are specified by temporal logic formulas. Different temporal logics exist, each with specific operators to explicitly reason about time or about precedence relationships between states. The temporal logics mostly used relate to the latter: the Linear Time Logic (LTL), in which time is considered linear, and the Computation Tree Logic (CTL) in which alternative time lines are considered (Clarke et al., 1999).

In the asynchronous dynamics of a logical model, a state may have multiple successors and hence lead to alternative paths, which makes CTL particularly useful. To check reachability properties, as illustrated in Section 5.2, we use CTL temporal operators, with the following syntax and semantics (see Clarke et al., 1999 for a complete reference of CTL operators):

- $EF(\Phi)$, there is at least one path leading to a state satisfying the property Φ ;
- $E[\Psi U \Phi]$, there is at least one path satisfying Ψ until it reaches a state satisfying Φ .

In the verification of software/hardware systems, a property is true if and only if it is true for every state in the set of initial states. However, when verifying biological systems, one is often interested in the existence of a reachability path from at least one of the initial states. The solution lies in the specification of the negated property (i.e., absence of reachability), which forces the model checker to answer false if there is at least one reachability path (used in Section 5.2).

A popular model checker is NuSMV (Cimatti et al., 2002). GINsim provides an export into a NuSMV description with the model rules, updating scheme and a (set of) initial state(s), together with other optional parameters. In a NuSMV description, the (set of) initial state(s) is specified using the keyword `INIT`, and the (set of) properties is specified using the keyword `SPEC` (cf. Section 5).

3.2. Assessing Model Behaviors upon Input Variations

Recall that input components have no associated regulatory function and are thus generally kept constant throughout simulation. This means that there are no transitions between states of the STG differing on values of input components (see Figure 2). However, these disconnected STG sub-graphs can be

connected by adding bi-directional transitions, which account for unconstrained variations of the input components. Using model checking tools, it is then possible to check properties for which inputs freely vary along a simulation. In order to account for a distinct semantics of inputs and internal (regulated) components, the Action Restricted Computation Tree Logic (ARCTL) is used (Lomuscio et al., 2007; Monteiro and Chaouiya, 2012). ARCTL extends CTL, imposing an additional path restriction on a subset of inputs while letting the remaining inputs to freely vary. This temporal logic was implemented in NuSMV-ARCTL, which extends NUSMV. In Section 5.1, we take advantage of a subset of ARCTL operators with the following syntax and semantics (see Lomuscio et al., 2007 for a complete description of ARCTL operators):

- $EAF(\alpha)(\Phi)$, there is at least one path leading to a state satisfying Φ , and the input restriction α must be satisfied along this path;
- $AAG(\alpha)(\Phi)$, all the states of all paths must satisfy Φ , and the input restriction α must be satisfied along these paths.

Other approaches have been developed to simulate Boolean models under stochastic and continuous environments (Helikar and Rogers, 2009; Helikar et al., 2012). Considering a synchronous update, a model input can be allocated a probability to be in its active state at each simulation step (see Figure 2D). This probability may represent finer levels of external signals. Furthermore, once a Boolean network has reached an attractor, the average active/inactive states of each component over the entire attractor can be calculated providing a characterization of the component activity level in this attractor (Todd and Helikar, 2012). Varying continuously the probabilities of input states (from 0 to 1), input-output dose-response (titration) curves can be generated, similar to those traditionally produced in experimental studies, for example to study the effects of different concentrations of a drug (Madrahimov et al., 2013) or of different concentrations of receptor ligands as in Figure 6 (Helikar et al., 2013a). Currently, this approach is supported by the Cell Collective (cf. Section 2.3), and by the stand-alone command-line simulation engine, ChemChains (Helikar and Rogers, 2009).

3.3. Model Reduction

A natural solution to lessen the combinatorial explosion issue is to reduce the size of the model. Any reduction potentially alters the properties of a model by modifying its dynamics. However, when the reduction impacts on the dynamics are well mastered, the analysis of a reduced model can be used to deduce interesting properties of the original model. This is the case of the reduction method that removes components while properly modifying the logical functions of their targets, which thus become directly affected by the regulators of the removed components (Naldi et al., 2011, 2012; Saadatpour et al., 2013). As a consequence, a self-regulated component cannot be removed, firstly because this definition is not applicable, but also because regulatory circuits are known to drive important dynamical properties and thus should not be concealed (Thieffry, 2007). The key point about this reduction is that it does not

generate novel trajectories and thus reachability properties that are verified in the reduced model are also true in the original model. Furthermore, Naldi et al. (2011) demonstrated that all the stable states and elementary cyclic attractors of the asynchronous dynamics are preserved. Because transitions of the original STG may be discarded (removing a component amounts to consider that its evolution is faster than that of the concurrent components), more complex attractors may be split in two or more complex attractors, while transient SCC may become terminal. However, Saadatpour et al. (2013) showed that, for constant input values, all the attractors are preserved when reducing input and pseudo-input components (i.e., components that are only regulated by inputs or by pseudo-inputs), as well as mediator components (which are characterized by a unique regulator and a unique target). Furthermore, both attractors and reachability properties are preserved when reducing output and pseudo-output components (i.e., components with no target, or whose targets are only outputs or pseudo-outputs, see Naldi et al., 2012).

3.4. Perturbation Analyses

In the logical framework, it is straightforward to define perturbations. Perturbations affecting model components often merely amount to force the corresponding variables to take specific values. For example, to specify a knock-out, it suffices to set the variable to 0, whereas for an ectopic expression the variable is set to its maximal value. Stimulation of a signaling pathway at the receptor level can be simulated by setting the variable describing the receptor to 1, and blockage with a drug of a protein by setting it to 0. By modifying the regulatory functions, subtler perturbations can be defined as, for example mutations in a promoter region, turning a component insensitive to a given regulator (cf. Section 5.2).

4. MODEL AND DATA INTEGRATION

4.1. Integration of Experimental Data

Because logical models provide a flexible framework to encode different biological events, with various granularities, they are particularly well suited to examine experimental data. Perturbations (genetic alterations, treatment with drugs or ligands, etc.), can be easily encoded in the model (cf. Section 3.4), and simulation results can then be mapped to the measured values of specific biological components upon these perturbations.

Different types of experimental data have been integrated within logical models. Genetic data are commonly used to define models and simulations (e.g., mutations or knockdown conditions), for various model organisms, from microbes (Thieffry and Thomas, 1995), to cancer (Remy et al., 2015), and many others. The data type used as readout depends on the system under study. In the case of gene regulatory networks, gene expression data are typically used, while for signal transduction, protein phosphorylation data are normally used. It is also possible to include non-molecular data, such as phenotypic measurements like growth, which is useful e.g., to connect the

effect of a drug on a signaling pathway with its effect on cellular growth (Kirouac et al., 2013; Flobak et al., 2015).

The integration of experiments and model can occur at different levels: (i) *a priori* in the building phase, to define or refine the model, (ii) *a posteriori*, to fit a generic model and obtain a model specific to certain conditions, and (iii) to (in)validate a model by challenging it to predict experimental data under specific conditions.

Model fitting to data allows to refine a given model structure relying on dedicated experiments. Because general network information is often not cell or context specific, such refinements lead to models that describe more accurately specific cellular situations. Such model adjustment can be done manually, by iteratively changing the model and testing how well the resulting model matches experimental data. For high-throughput data sets, this process has been automatized by casting it as an optimization problem (Saez-Rodriguez et al., 2009). This methodology can be applied in multiple biological contexts and to different types of data. In the case of signaling, as stated above, proteomic data are particularly adequate and can be obtained with antibody based platforms, such as protein arrays or luminex (Saez-Rodriguez et al., 2009), or using mass spectrometry (Terfve et al., 2015). Gene expression data can also be used (Crespo et al., 2013; Keller et al., 2016).

In addition to using experimental data for logical model construction, various types of data available in many databases can be exploited, in turn, to interpret simulation experiments and further validate the models and associated predictions. The advantage of dynamical models is that they can generate hypotheses about any targeted component, or about the system as a whole. For example, Puniya et al. (2016) interrogated a comprehensive signal transduction network model under all possible knock-in and knock-out perturbations, resulting in the identification and ranking of the most and least influential model components. These components were further mapped on various databases, resulting in the prediction of a new combinatorial drug target in a cancer setting.

In practical terms, standardized names and proper annotations using controlled vocabularies are essential for a correct integration of models and data. This issue is discussed in the next section.

4.2. Exchange Formats and Model Documentation

As the popularity of logical modeling increases, standardization issues have to be tackled. To this intent, the informal consortium CoLoMoTo (<http://www.colomoto.org>) gathers researchers developing logical models, methods and tools (Naldi et al., 2015). The definition of a common file format was identified as a primary requirement to allow model exchange and software interoperability. Model encoding in a standard format facilitates model reuse for extension or composition. In the context of SBML Level 3 (Systems Biology Markup Language Hucka et al., 2003), the SBML Qual (for qualitative) package has been defined to store logical models (Chaouiya et al., 2013, 2015). This format is currently supported by a number of software tools, including

GINsim, Cell Collective, CellNOpt, the tools mentioned in this paper. Hence, models stored in the SBML qual format can be exchanged between these tools. Thanks to the LogicalModel library, GINsim also provides an export of Boolean models to MaBOSS, in addition to several other formats (Chaouiya et al., 2013).

To allow reproducibility of *in silico* experiments, simulation settings must be specified along with the model itself. These settings include the initial condition(s) and a precise description of the updating scheme. Model perturbations may be also considered as specific simulation settings. For example, the software GINsim allows to store all this information in the form of a set of parameter settings (or simulation scenarios). Cell Collective stores simulation settings in a database. The Simulation Experiment Description Markup Language (SED-ML) has been defined as a standard format for encoding simulation experiments (Waltemath et al., 2011). One objective of CoLoMoTo is to promote the use of such a format, possibly by extending it to support specificities of logical modeling.

Proper documentation and annotation are crucial for reuse and expansion of computational models by the community. Often, published models lack information (evidence and/or clear assumptions) documenting model components, interactions and rules. Several efforts already exist to address this issue. The Minimum Information Requested In the Annotation of biochemical Models (MIRIAM) (Le Novère et al., 2005) was developed to standardize the type of information (e.g., connections to controlled vocabularies as well as to various databases) that should be included as model metadata. While MIRIAM (and other standards) provides minimal guidelines to ensure model reproducibility, additional efforts are needed to increase the overall quality (breadth and detail) of model documentation. For instance, modelers and curators can provide detailed and exhaustive evidences supporting model components and interactions when available, or assumptions in the case of unavailable experimental observations. This is facilitated in Cell Collective, which provides a Knowledge Base for each model.

Finally, BioModels database (Chelliah et al., 2015) and other model repositories such as those of GINsim and Cell Collective are also essential to ensure that models are available to the community for reproducibility of the results as well as for model reuse.

5. LOGICAL MODELING AND ANALYSES OF TWO DISTINCTIVE APPLICATIONS

5.1. Application 1: T Cell Signaling

T lymphocytes play a central role in the adaptive immune response in mammals. Cytotoxic CD8⁺ T cells kill cells infected by viruses or malignant cells, whereas CD4⁺ T helper (Th) cells orchestrate the function of a large diversity of effector immune cells (including B cells, macrophages, granulocytes, and NK cells) (Murphy et al., 2012). Activation of T cells and their subsequent differentiation into effector or regulatory cells result from the integration of a large panel of signals from their microenvironment. Initially in a naïve state, T cells are

activated by three main types of signals: (i) T cell receptor (TCR) activation, through the specific recognition of foreign antigens presented by antigen presenting cells (APCs), (ii) co-inhibitory and co-stimulatory signals, and (iii) cytokines. The integration of these multiple signals initiates a plethora of signaling cascades, regulating complex and intertwined networks, which ultimately control T cell activation, proliferation and differentiation into effector or regulatory cells expressing specific markers.

For example, Th1 subtype is characterized by the production of interferon gamma (IFN- γ), leading to the clearance of intracellular pathogens, whereas Th2 cells secrete the cytokines interleukin-4 (IL-4), IL-5 and IL-13, involved in the elimination of helminths. Recently, additional Th subsets (e.g., Th17, Treg, Tfh, Th9, Th22) have been characterized. Furthermore, recent experimental evidences emphasize the diversity and plasticity of T cells, challenging the classical picture of irreversible branching differentiation (Nakayamada et al., 2012).

In order to decipher the mechanisms underlying T lymphocyte activation and differentiation, various logical models have been proposed, each addressing specific aspects (cf. **Table 1**). Hereafter, we discuss a sample of these modeling efforts to emphasize specific aspects of modeling and analysis, as well as insights into the regulation of T cell activation and differentiation.

Relying on the initial identification of Th1 and Th2 dichotomy, Mendoza (2006) proposed a logical model of the differentiation network accounting for some aspects of Th commitment toward these two cell types. The author could capture Th1 and Th2 cellular types in terms of stable states of the model, and got further insights into the intracellular circuits involved in the delineation of the corresponding basins of attractions. Naldi et al. (2010) extended Mendoza's model to cover additional signaling pathways and Th subsets (Th17, Treg), using GINsim for model construction and analysis. As the model was too large for a direct analysis of its dynamics, the authors applied a reduction method (cf. Section 3.3), which led to a model encompassing 34 components, amenable to analysis through systematic simulations. Following the identification of all the stable states, these were grouped according to relevant phenotypic Th markers, abstracting away input values. The model accounts for the canonical Th1, Th2, Th17, and Treg subtypes, as well as for additional hybrid Th subtypes coexpressing combinations of canonical Th markers. Finally, the authors assessed the stability of the identified Th subtypes, under specific polarizing environmental conditions (defined by model input values), by iterating rounds of simulation of the reduced model dynamics. Interestingly, this reachability analysis emphasized the plasticity of the Th subtypes upon environmental changes, with some cell types predicted to be highly labile (Th17, Treg) whereas other are shown to be more robust (Th1, Th2).

Extending this model, Abou-Jaoudé et al. (2015) proposed a multi-valued model accounting for novel canonical Th subtypes, namely Th9, Th22, Tfh, with the integration of additional transcription factors (e.g., PU.1, Bcl6) and cytokine pathways involved in Th cell commitment. Following the approach of Naldi et al. (2010) and considering a reduced version of the model (cf. **Figure 3**), all the stable states were identified and

TABLE 1 | Selected logical models of T cell signal transduction and gene regulation.

| Publication | Characteristics (nb components) | Dynamics | Availability |
|---|---|--|--|
| Mendoza, 2006 | Multilevel model; CD4+ T cell differentiation regulatory network (17) | Stable state analysis, perturbation analysis, circuit analysis | GINsim |
| Saez-Rodríguez et al., 2007 | Boolean model; T cell receptor signal transduction network (94) | Stable state analysis, input and perturbation analysis | Cell Collective |
| Zhang et al., 2008; Saadatpour et al., 2011 | Boolean model; T cell survival signal transduction network (60) | Asynchronous update, perturbation analysis, model reduction, attractor identification | GINsim Cell Collective |
| Naldi et al., 2010 | Multilevel model; CD4+ T cell differentiation regulatory network (65) | Asynchronous update, model reduction, circuit analysis, Th cell plasticity | GINsim |
| Martínez-Sosa and Mendoza, 2013 | Boolean model; CD4+ and CD8+ T cells regulatory network (50) | Synchronous update, attractor analysis, perturbation analysis | Cell Collective |
| Miskov-Zivanov et al., 2013 | Multilevel model; TCR signaling pathways (38) | Random asynchronous update, introduction of delays, duration of input stimuli modeled as a number of updating rounds | Cell Collective |
| Conroy et al., 2014 | Boolean model; TCR and integrin signaling network and T cell differentiation regulatory network (188) | Synchronous update, stochastic inputs, perturbation effects on downstream components | Cell Collective |
| Oyeyemi et al., 2015 | Boolean model; HIV-T cell interaction network (137) | Stable state analysis, perturbation analysis | Cell Collective |
| Abou-Jaoudé et al., 2015 | Multilevel model; CD4+ T cell differentiation regulatory network (101) | Asynchronous update, model reduction, stable state analysis, model checking, Th cell plasticity | GINsim |
| Martínez-Sánchez et al., 2015 | Boolean model; CD4+ T cell differentiation regulatory network (85) | Model reduction, attractor analysis, perturbation analysis, Th cell plasticity | Cell Collective BioModels DB (non-curated branch) |

grouped according to phenotypic markers, thereby defining expression patterns associated with each canonical Th subtype. This analysis allowed to capture the novel canonical subtypes and predicted hybrid subtypes in terms of stable states. Noteworthy, the interpretation of the input dependency of the stability of these states is hindered by the gigantic number of input configurations (2^{21} value combinations of the 21 binary inputs). To cope with this combinatorial explosion, one can further cluster these stable states according to relevant input signatures.

Abou-Jaoudé et al. (2015) used model checking to efficiently analyze Th cell plasticity under relevant polarizing conditions. More precisely, using NuSMV-ARCTL (cf. Section 3.2), reachability properties between the canonical Th subtypes were systematically analyzed, considering relevant cytokinetic environmental conditions. The following generic ARCTL property was specified to verify the existence of a reachability path from a canonical Th pattern c_1 toward a (stable) canonical Th pattern c_2 under an input condition e (the & operator denotes the conjunction):

$$\text{INIT } c_1; \text{SPEC } \text{EAF}(e)(c_2 \ \& \ \text{AAG}(e)(c_2)).$$

Results were synthetically represented in the form of a reprogramming graph, which reproduces various polarizing events experimentally observed and uncovers many reprogramming scenarios between Th subtypes (see **Figure 4**). In particular, several strategies allowing Th1 vs. Th2 interconversions could be identified, in accordance with recent experimental observations challenging Th1-Th2 dichotomy (Antebi et al., 2013).

Other scenarios where a Th subtype can follow distinct fates under the same environmental conditions were also unraveled by this analysis. To get comprehensive insights into the alternative trajectories underlying different cell decisions, a HTG representation of the dynamics can be used. **Figure 5** provides an example of such a representation starting from Th22 cells and immersing them into a Treg polarizing environmental condition. We see here that three stable states can be reached, one corresponding to a Th17 cell type, and two corresponding to Treg cell types. The cell decision between these phenotypes mainly depends on the concurrent activation of Ror γ t (the master regulator of Th17 cells) and Foxp3 (the master regulator of Treg cells). Further insight into the reachability of the three attractors can be extracted by performing a reachability analysis with the



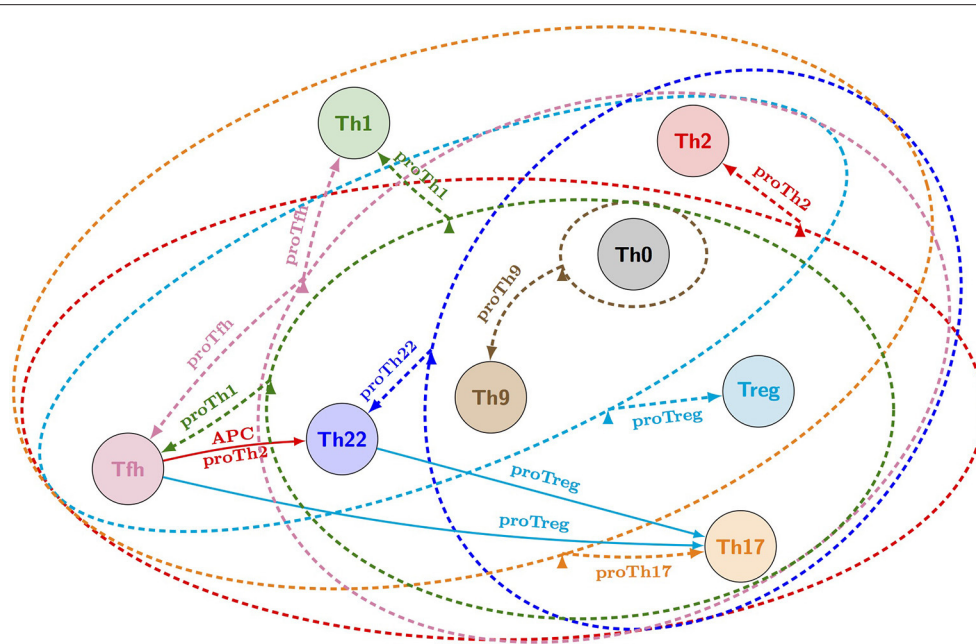


FIGURE 4 | Reprogramming graph considering the canonical Th subtypes (generated with the model checker NuSMV-ARCTL; adapted from Abou-Jaoudé et al., 2015). Ellipses gather all subtypes that, under the same environmental condition, differentiate toward a particular stable subtype (defined in Table 3 in Abou-Jaoudé et al., 2015). Dashed arrows connect ellipses to a (set of) differentiated state(s) and are labeled with the corresponding environmental conditions. Solid arrows denote specific reachability conditions between pairs of subtypes, under a particular environmental condition. Colors of arrows and ellipses indicate the environmental conditions of the corresponding subtype color. For example: from Th2, Th22, Th9, Th0, Treg, and Th17 subtypes (gathered in the pink ellipse), a “proTfh” condition leads to reprogramming into both Tfh (pink node) and Th1 subtypes; while from Th22, a “proTreg” condition leads to reprogramming into both Th17 and Treg subtypes.

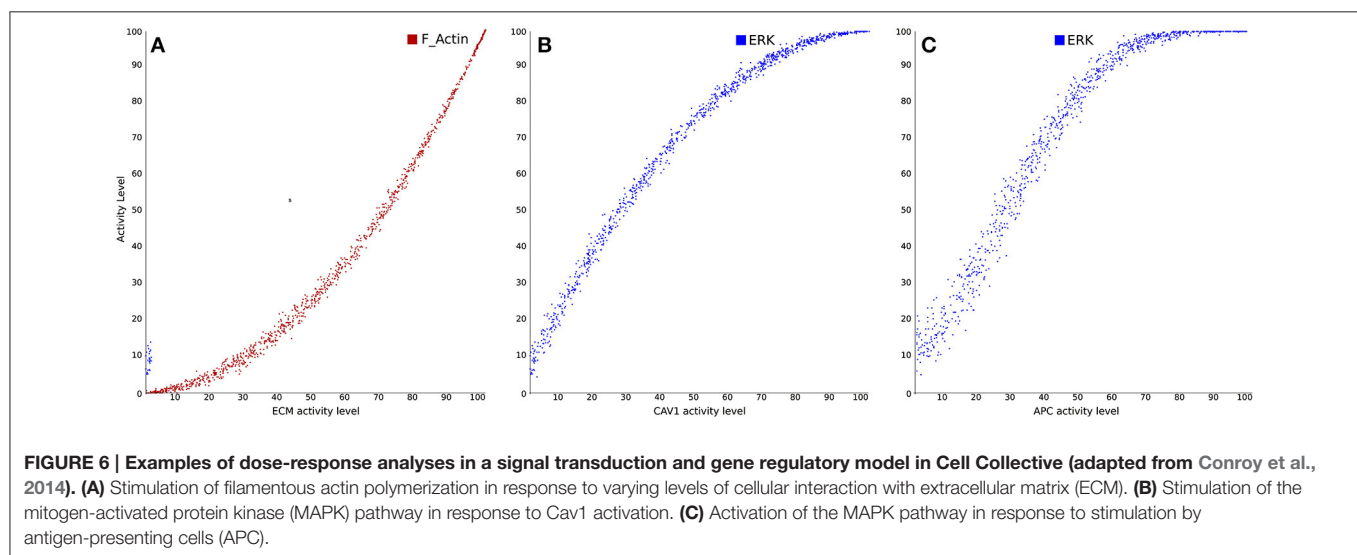
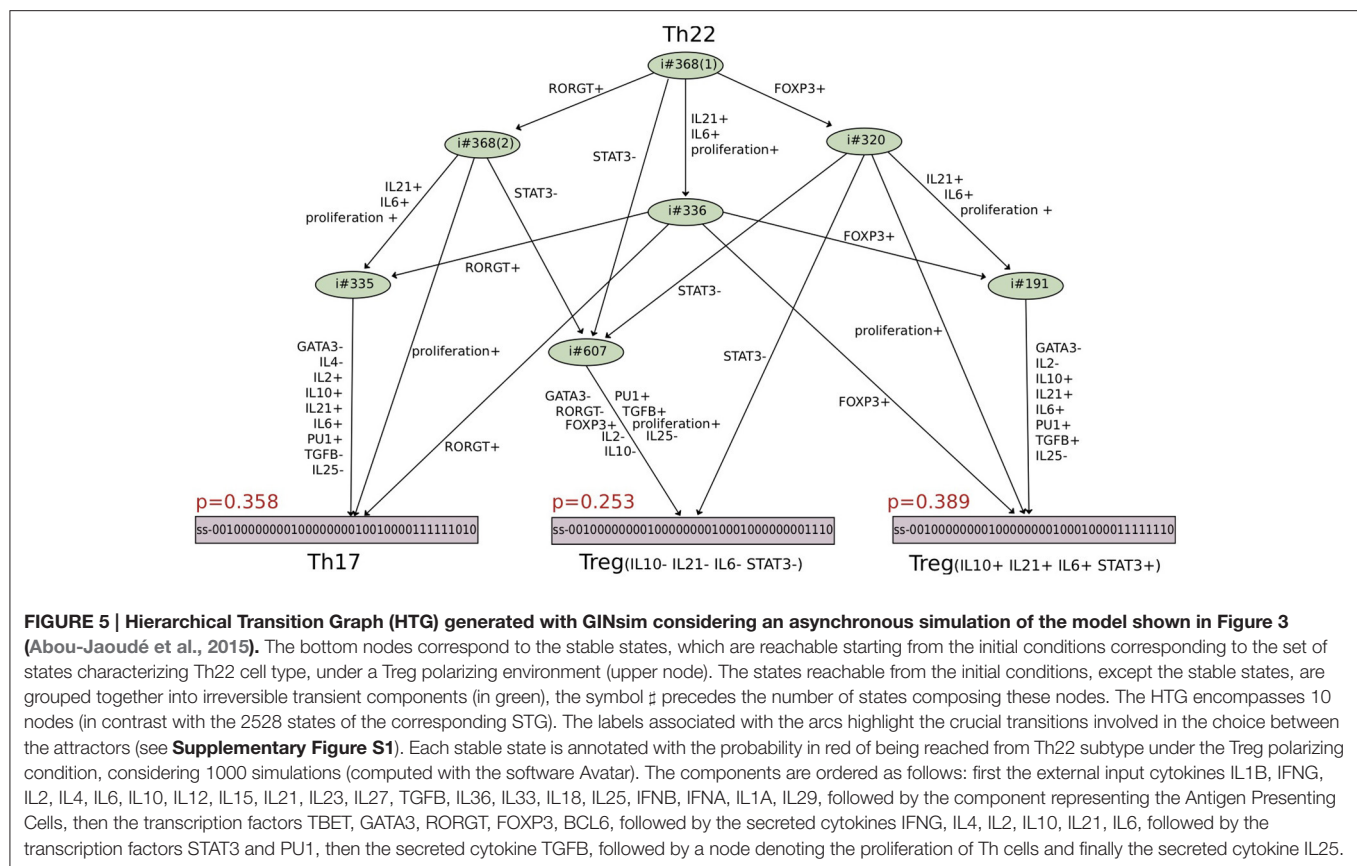
Avatar algorithm, quantifying the reachability probability of each attractor (Mendes et al., 2014). A thousand Avatar simulations were enough to observe a stabilization of the reachability probabilities of the three stable states. These indicated a higher probability to reach the Treg stable states (0.642) than the Th17 state (0.358), suggesting that a Treg environment would favor Th22 cells reprogramming toward a Treg rather than a Th17 phenotype (Figure 5).

Other modeling works have focused on the signaling pathways underlying T cell activation, survival and proliferation. Saez-Rodriguez et al. (2007) established a Boolean model of T cell activation following the engagement of TCR and co-stimulatory receptor CD4 and CD28, using CellNetAnalyzer for model definition and analysis. Here, an analysis based on steady-state approximation was used. The reasoning being that in signal transduction several different time scales operate; a first wave of activation occurs upon stimulation with ligands and drugs, which often takes only a few minutes, and this is followed by feedback processes, which are typically slower. This approximation is clearly not accurate, but it permits the consideration of large networks in a simple and efficient manner. The model was able to recapitulate a large number of published data in both wild-type and knock-out conditions, as well as to predict unexpected signaling patterns after specific stimulation of the co-receptor CD28 and knock-out of the kinase Fyn, which were subsequently experimentally validated (Saez-Rodriguez et al., 2007).

Finally, several logical models were proposed to analyze T cell signaling networks in pathogenic situations, in particular in the context of T cell leukemia, a disease characterized by an abnormal proliferation of T cells (Zhang et al., 2008; Saadatpour et al., 2011; Conroy et al., 2014). Specifically, Conroy et al. (2014) developed a logical model to better understand the role of caveolin-1 (Cav1; an important regulator of endocytosis) in T-cell leukemia. Figure 6 illustrates input-output simulations and analyses demonstrating the ability of the model to correctly reproduce previously described and documented relationships between different components of the modeled network. Besides, the model allowed to identify the protein products most affected by CAV1+/+, CAV1+/-, and CAV1-/- under immunocompetent and immunocompromised conditions. Simulation results suggested that CAV1 expression regulates Ras-related C3 botulinum toxin substrate 1 (RAC1), B-cell lymphoma/leukemia 10 (BCL10), GATA-binding protein 3 (GATA3), CD26, and CD28. In addition to validating these predictions in Cav1 knock-out mice, model results were further successfully validated against gene expression signatures obtained from the Gene Expression Omnibus (GEO) database.

5.2. Application 2: Cell Cycle Control

Tightly controlled by a sophisticated regulatory network involving transcriptional regulations and protein modifications, cell proliferation involves successive phases governing genome replication (S phase) and cell division (mitosis or M phase),



separated by regulated irreversible transitions (checkpoints). The main components and regulatory interactions controlling cell cycle were initially identified in simplified model systems, including fission and budding yeasts, as well as early *Xenopus* zygotic mitoses. The underlying core networks have been modeled using differential equations, leading to novel insights into their organization and dynamical properties (see Ferrell

et al., 2011; Tyson and Novák, 2015 for recent reviews). However, extension and analysis of such differential models become really difficult as the number of experimentally identified components and interactions increases. This led several groups to consider Boolean or more sophisticated logical formalisms to build comprehensive models of cell cycle control networks (**Table 2**). Cell cycle networks present particular difficulties from the point

TABLE 2 | Selected logical models of cell cycle networks in different organisms.

| Publication | Organism | Characteristics (nb components) | Dynamics | Availability |
|------------------------------|---------------|---|---|---------------------------|
| Li et al., 2004 | Budding Yeast | Boolean model; Threshold logical functions (11) | Synchronous update, G1 stable state attracting most trajectories | GINsim (adapted model) |
| Fauré et al., 2006 | Mammals | Boolean model; Regulatory graph and standard logical functions (10) | Synchronous, asynchronous and mixed updating scheme; cyclic attractor plus quiescent stable state | GINsim Cell Collective |
| Davidich and Bornholdt, 2008 | Fission Yeast | Boolean model; Threshold logical functions (10) | Synchronous update, G1 stable state attracting most trajectories | GINsim (adapted model) |
| Irons, 2009 | Budding Yeast | Boolean model; Regulatory graph with 4 phenomenological nodes and standard logical functions (18) | Synchronous and temporized updating schemes, single cyclic attractor | GINsim Cell Collective |
| Fauré et al., 2009 | Budding Yeast | Multilevel model; Regulatory graph and standard logical functions (32) | Priority classes, single cyclic attractor | GINsim |
| Sahin et al., 2009 | Human | Boolean model; Regulatory graph and standard logical functions (20) | Asynchronous update, 3 stable states, transient oscillations | GINsim Cell Collective |
| Todd and Helikar, 2012 | Budding Yeast | Boolean model; based on Irons (2009)'s model (20) | Analysis over variation of inputs, which are allocated probabilities to be active | Cell Collective |
| Flobak et al., 2015 | Human | Boolean model; Regulatory graph and standard logical functions, no input (77) | Asynchronous update on a reduced model, a single stable state denoting cell proliferation | GINsim |

of view of logical modeling. On the one hand, cell cycling behavior tentatively corresponds to a cyclic attractor, or at least to some multiple state pathway in the STG (rather than to a logical stable state as for the Th subtypes mentioned above), which are hard to compute. On the other hand, of most importance is the precise succession of component switches along the cell cycle, ensuring the proper temporal articulation of the molecular processes required for successful genome replication and repartition, along with timely and balanced cell division.

The studies listed in **Table 2** rely on different modeling assumptions (e.g., using generic or specific rules, and considering specific updating schemes). By and large, relying on qualitative information, the authors were able to capture the succession of key events involved in cell cycle. Moreover, several studies recapitulate the effect of various kinds of perturbations (losses- or gains-of-function, see e.g., Fauré et al., 2006; Fauré et al., 2009; Irons, 2009). Fauré and Thieffry (2009) published a comparative review of cell cycle logical models (predating 2009). An interesting observation was the conservation of a functional negative regulatory circuit at the core of the cell cycle engine, involving cyclin B and Cdc20 (or their orthologs in other species), as well as of several coupled positive regulatory circuits. Here, we restrict ourselves to a few studies in order to emphasize specific aspects of logical modeling analyses.

Based on the differential model proposed by Novák and Tyson (2004) and Fauré et al. (2006) defined a Boolean model for the core network driving the entry of mammalian cells into cell cycle. This model accounts for the existence of a quiescent

stable state (in the absence of growth factors, represented by the shutoff of cyclin D, the input component), as well as for a cyclic attractor characterized by the periodic activities of the cyclins A, B and E, which drive the cell cycle through key transitions by enabling the phosphorylation of a number of substrates by their catalytic partners, the cyclin-dependent kinases (CDKs). This model further includes the three main inhibitors of the cell cycle: the retinoblastoma protein Rb, the CDK inhibitor p27/Kip1 and the proteasome complex represented by its two co-activators Cdh1 and Cdc20. Finally, this model accounts for the role of the E2 ubiquitin conjugating enzyme UbcH10, which participates in Cdh1 dependent degradation of cyclin A. This extension of the original differential model explains how the auto-ubiquitination of UbcH10 probably prevents cyclin A from degradation by the APC in G1 phase. Complex formation and protein sequestration were modeled in terms of logical rules associated with the target proteins, which enabled the author to keep the number of components considered to the low end (ten components). Although very simplified, this model broadly reproduced the sequence of molecular events along the normal cell cycle, for both synchronous and asynchronous updating schemes. The authors further considered a list of documented perturbations to validate their model. Although the simulations of various perturbations were shown to match experimental observations, it was not the case for some documented perturbations, including for a knock-out of cyclin E.

Traynard et al. (2015) revisited this model to solve the remaining discrepancies in the light of recent data (see **Figure 7**). As hinted already in the seminal study by Fauré et al. (2006),

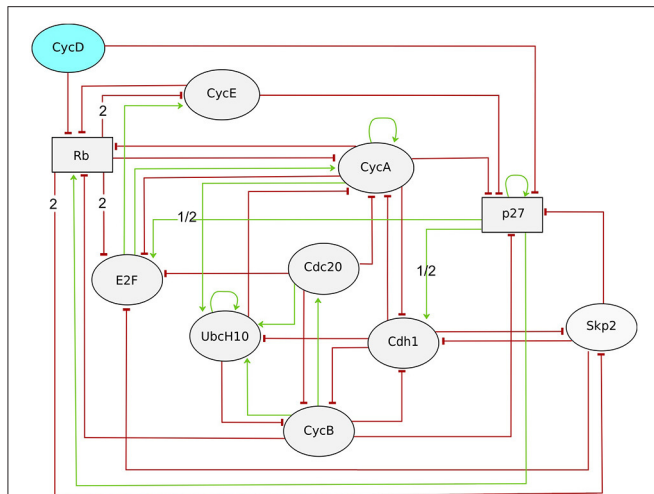


FIGURE 7 | Regulatory graph of the mammalian cell cycle model (Traynard et al., 2015). The input node, CycD accounts for the positive signal, as Cyclin D is activated by growth factors. All components are Boolean, except Rb and p27 (see Text). Interactions requiring the higher threshold (value 2) or having different effect depending on the threshold value (1/2) are labeled accordingly.

the authors considered the use of a ternary variable for the cell cycle inhibitor Rb, which can be phosphorylated at multiple sites, associated with different activities. Similarly, they associated a ternary variable with p27 to account for its significant but incomplete degradation in the presence of CycD and in the absence of CycA and CycE. They further included the F-box protein Skp2 in the model. Skp2 promotes the degradation of phosphorylated p27 and thereby enables its degradation. Skp2 degradation is promoted by Rb binding to Cdh1. Skp2 thus links the two cell cycle repressors Rb and p27, and provides an additional mechanism by which Rb can arrest the cell cycle. In order to assess the benefits of each modification, model checking was used to verify the existence (or the absence) of specific trajectories characteristic of the cell cycle dynamics. More specifically, a generic CTL temporal logical formula (see Section 3.1) was used to verify the existence of a trajectory complying with a sequence $S_1, S_2, S_3, \dots, S_{n-1}, S_n$, each denoting a set of states defined by constraints on some of the model components:

$$\text{INIT } S_1; \text{ SPEC } !E[(S_1) \cup (S_2 \& E[(S_2) \cup (S_3 \& \dots E[(S_{n-1}) \cup (S_n)]))]]].$$

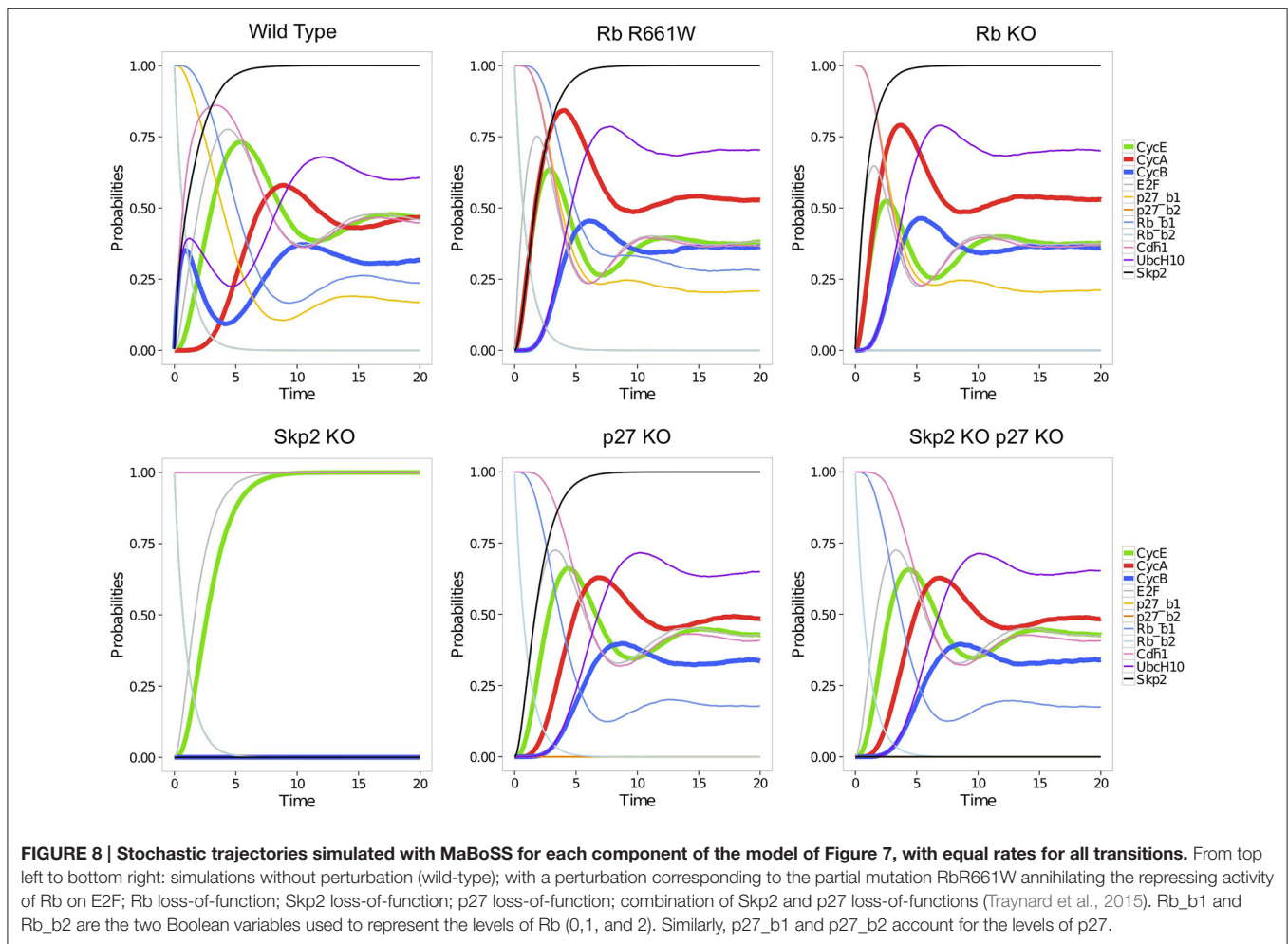
Here, the negation (denoted by the operator $!$) is used to obtain a counter-example, from the model checker, whenever the property is false, containing the desired trajectory complying with a sequence $S_1, S_2, S_3, \dots, S_{n-1}, S_n$. As a result, the authors obtained a generic multi-valued logical model of the mammalian cell cycle that qualitatively matches the most salient dynamical properties of the normal cell cycle, in particular at the G1/S transition, as well as the phenotypes of many mutants (Traynard et al., 2015).

More quantitative characterizations of asymptotic behaviors can be provided by stochastic simulations using MaBoSS (see Section 2.3). As MaBoSS is restricted to Boolean models, the ternary node Rb was split into two Boolean nodes Rb1 and Rb2, associated with the first and second Rb thresholds, respectively (and similarly for p27). The stochastic trajectories computed for this model reflect the kinetics of the cell cycle progression driven by the input cyclin D (see Figure 8). Transient oscillations can be observed as the trajectories all start in G0 (with Rb1, Rb2, p27, and Cdh1 the only active nodes) and progressively desynchronize. It is particularly interesting to compare the trajectories obtained for wild type (WT) vs. perturbed conditions. The trajectories obtained for five perturbations illustrate the role of Rb and of the pathway Rb-Skp2-p27 in the model (Figure 8). Two perturbations were considered for Rb: the full loss-of-function (Rb KO), and a partial loss-of-function, where Rb loses its ability to repress E2F, but conserves its repressing activity on Skp2 (Rb R661W). The resulting stochastic trajectories highlight the role of Rb in the sequential activation of cyclin E and cyclin A, ensured by the repressing activity of the two underphosphorylated forms of Rb on E2F: in the WT case, the activation of cyclin A is clearly delayed relatively to the activation of cyclin E. In contrast, in the absence of the repressing effect of Rb on E2F, cyclin E and cyclin A are activated at the same time. The lack of significant difference between the trajectories of Rb R661W and Rb KO suggests that the repression of Skp2 by Rb has no major impact on the cell cycle. However, this interaction is necessary to ensure the quiescent state in the absence of cyclin D. Skp2 loss-of-function (Skp2 KO) arrests the cell cycle (Figure 8), presumably due to the stabilization of p27. Indeed, the oscillations are restored in the double mutant Skp2 KO p27 KO.

In an independent study focusing on Yeast cell cycle control, Todd and Helikar (2012) built on the model of Irons (2009) and showed that cell phenotypes can be modeled as ergodic sets irreducible sets of states of the corresponding Markov chain; i.e., set of states that cannot be left once reached), by defining probabilities for the input components to be active and modeling these signals as continuous variables. In this work, the cell cycle was analyzed as a sequence of models, each accounting for a specific phase of the cycle, which allowed to characterize the (continuous) dynamics of all regulatory components along each phase, and more closely compare them to various experimental observations. Modeling extracellular signals as continuous variables (i.e., cell size) resulted in the finding that the yeast cell cycle network is stable under different patterns of cell growth. That is, as long as the checkpoints are appropriately activated (i.e., the environment is stable enough for the successful completion of the current phase), the modeled cell progresses through the cycle, independently of its size. Furthermore, the continuous dynamics of the model components were found consistent with various experimental studies.

6. DISCUSSION

After introducing the logical modeling framework and a range of methodological advances to analyze dynamical properties of



these discrete models, we have presented a number of assets of this approach through two important case studies. Here, we discuss further issues and complementary approaches.

Besides the consideration of probabilistic input values, we have focused on non-stochastic models (recall that asynchronous dynamics is non-deterministic but not random). However, several methods have been proposed to include noise in Boolean models. For example, accounting for uncertainty in the regulatory functions, Shmulevich et al. (2002) associate each component with a set of regulatory functions, one being randomly selected at each step of the simulation. Another option consists in randomly taking the complements of the regulatory function outcomes (Alvarez-Buylla et al., 2008). In Garg et al. (2009), the authors consider potential failures of the regulatory functions. For all these stochastic variants, the synchronous scheme was adopted.

It is worth mentioning that several continuous transpositions of logical models have been proposed, for example considering fuzzy logic (Aldridge et al., 2009; Morris et al., 2011), or transforming Boolean models into ordinary differential equations (Mendoza and Xenarios, 2006; Wittmann et al., 2009). The reverse transformation has been formally addressed for the

specific class of piecewise affine differential models (Batt et al., 2008; Chaves et al., 2010).

As shown in Section 3 with the usage of model checking, logical models are amenable to sophisticated formal methods. Initially developed for software and hardware systems, these techniques are indeed well adapted for logical model identification (e.g., constraint programming, see Corblin et al., 2010 and Answer Set Programming, see Videla et al., 2015) and for model analysis (e.g., satisfiability problem (SAT) for the identification of the attractors of Boolean models, see Dubrova and Teslenko, 2011).

Although progress has been made with the definition of SBML qual, the SBML Level 3 Qualitative Models Package (Chaouiya et al., 2013), further efforts are needed to ensure model exchange, reuse and extension. A first issue concerns reproducibility of modeling studies. This can be achieved first by providing model files, in BioModels database (Chelliah et al., 2015), or in model repositories such as those provided by Cell Collective or GINSim (see Section 2.3). Second, modeling assumptions and simulation settings should be precisely described. For example, we have underlined that model properties can vary depending on the adopted updating scheme (Section 2.2). Furthermore, model

extensions often simply refer to the addition of components, but it can also consist in refining the model with a stochastic extension (e.g., with probabilistic input values as in **Figure 6**). The different formalism extensions evoked above together with many others still need to be precisely characterized, managed within a control vocabulary and supported in a future SBML qual version. Further integration with core SBML Level 3 concepts will be needed to support the encoding of hybrid models combining features of both discrete and continuous formalisms. It is the purpose of CoLoMoTo (the Consortium for Logical Models and Tools) to stimulate and coordinate such developments.

AUTHOR CONTRIBUTIONS

WA, PT, and PM equally contributed to the manuscript, which content and organization have been designed by CC, DT, TH, and JS. CC coordinated the writing of the manuscript, and contributed particularly to Sections 2, 3. PM contributed to Section 3 in particular. JSR contributed to Section 4 and revised the manuscript. DT supervised the writing of Section 5 and particularly contributed to Sections 1, 5, and 6. WA contributed to Section 5.1 and PT to Section 5.2. TH contributed to Sections 4, 5. All authors have read and approved the final manuscript.

REFERENCES

- Abou-Jaoudé, W., Monteiro, P. T., Naldi, A., Grandclaudeon, M., Soumelis, V., Chaouiya, C., et al. (2015). Model checking to assess T-helper cell plasticity. *Front. Bioeng. Biotechnol.* 2(Suppl. 1):86. doi: 10.3389/fbioe.2014.00086
- Albert, R., and Thakar, J. (2014). Boolean modeling: a logic-based dynamic approach for understanding signaling and regulatory networks and for making useful predictions. *Wiley Interdisc. Rev. Syst. Biol. Med.* 6, 353–369. doi: 10.1002/wsbm.1273
- Aldridge, B. B., Saez-Rodriguez, J., Muhlich, J. L., Sorger, P. K., and Lauffenburger, D. A. (2009). Fuzzy logic analysis of kinase pathway crosstalk in tnfr/egf/insulin-induced signaling. *PLoS Comput. Biol.* 5:e1000340. doi: 10.1371/journal.pcbi.1000340
- Alvarez-Buylla, E. R., Chaos, A., Aldana, M., Benítez, M., Cortes-Poza, Y., Espinosa-Soto, C., et al. (2008). Floral morphogenesis: stochastic explorations of a gene network epigenetic landscape. *PLoS ONE* 3:e3626. doi: 10.1371/journal.pone.0003626
- Antebi, Y. E., Reich-Zeliger, S., Hart, Y., Mayo, A., Eizenberg, I., Rimer, J., et al. (2013). Mapping differentiation under mixed culture conditions reveals a tunable continuum of T cell fates. *PLoS Biol.* 11:e1001616. doi: 10.1371/journal.pbio.1001616
- Arellano, G., Argil, J., Azpeitia, E., Benítez, M., Carrillo, M., Góngora, P., et al. (2011). “Antelope”: a hybrid-logic model checker for branching-time Boolean GRN analysis. *BMC Bioinformatics* 12:490. doi: 10.1186/1471-2105-12-490
- Batt, G., de Jong, H., Page, M., and Geiselman, J. (2008). Symbolic reachability analysis of genetic regulatory networks using discrete abstractions. *Automatica* 44, 982–989. doi: 10.1016/j.automatica.2007.08.004
- Batt, G., Ropers, D., de Jong, H., Geiselman, J., Mateescu, R., Page, M., et al. (2005). Validation of qualitative models of genetic regulatory networks by model checking: analysis of the nutritional stress response in *Escherichia coli*. *Bioinformatics* 21(Suppl. 1):i19–i28. doi: 10.1093/bioinformatics/bti1048

FUNDING

WA has been supported by postdoctoral grants from the LabEx MemoLife and from the Ecole Normale Supérieure. ENS team further acknowledges the support of the French *Plan Cancer* (2014–2017), in the context of the project entitled *Modeling cell communication networks in breast cancer - CoMET*. PM was supported by FCT grants UID/CEC/50021/2013 and IF/01333/2013. CC acknowledges the support of the Fundação Calouste Gulbenkian.

ACKNOWLEDGMENTS

The authors would like to thank all members of the CoLoMoTo consortium for inspiring discussions and fruitful collaborations.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: <http://journal.frontiersin.org/article/10.3389/fgene.2016.00094>

Supplementary Figure S1 | This figure provides an illustration of a simple multi-valued model and its asynchronous dynamics represented as a State Transition Graph and as a Hierarchical Transition Graph.

- Béranguier, D., Chaouiya, C., Monteiro, P. T., Naldi, A., Remy, E., Thieffry, D., et al. (2013). Dynamical modeling and analysis of large cellular regulatory networks. *Chaos* 23:025114. doi: 10.1063/1.4809783
- Bonzanni, N., Garg, A., Feenstra, K. A., Schütte, J., Kinston, S., Miranda-Saavedra, D., et al. (2013). Hard-wired heterogeneity in blood stem cells revealed using a dynamic regulatory network model. *Bioinformatics* 29, i80–i88. doi: 10.1093/bioinformatics/btt243
- Bornholdt, S. (2008). Boolean network models of cellular regulation: prospects and limitations. *J. R. Soc. Interface* 5(Suppl. 1), S85–S94. doi: 10.1098/rsif.2008.0132.focus
- Brim, L., Česka, M., and Šafránek, D. (2013). “Model checking of biological systems,” in *Formal Methods for Dynamical Systems, Volume 7938 of Lecture Notes in Computer Science*, eds M. Bernardo, E. Vink, A. Pierro, and H. Wiklicky (Bertinoro: Springer), 63–112.
- Chabrier, N., and Fages, F. (2003). “Symbolic model checking of biochemical networks,” in *Computational Methods in Systems Biology, Volume 2602 of Lecture Notes in Computer Science*, ed C. Priami (Rovereto: Springer), 149–162. doi: 10.1007/3-540-36481-1_13
- Chaouiya, C., Béranguier, D., Keating, S. M., Naldi, A., van Iersel, M. P., Rodriguez, N., et al. (2013). SBML qualitative models: a model representation format and infrastructure to foster interactions between qualitative modelling formalisms and tools. *BMC Syst. Biol.* 7:135. doi: 10.1186/1752-0509-7-135
- Chaouiya, C., Keating, S. M., Béranguier, D., Naldi, A., Thieffry, D., van Iersel, M. P., et al. (2015). The systems biology markup language (SBML) level 3 package: qualitative models, version 1, release 1. *J. Integr. Bioinform.* 12:270. doi: 10.2390/biecoll-jib-2015-270
- Chaouiya, C., Naldi, A., and Thieffry, D. (2012). Logical modelling of gene regulatory networks with ginsim. *Methods Mol. Biol.* 804, 463–479. doi: 10.1007/978-1-61779-361-5_23
- Chaves, M., Albert, R., and Sontag, E. D. (2005). Robustness and fragility of Boolean models for genetic regulatory networks. *J. Theor. Biol.* 235, 431–449. doi: 10.1016/j.jtbi.2005.01.023

- Chaves, M., Tournier, L., and Gouzé, J.-L. (2010). Comparing Boolean and piecewise affine differential models for genetic networks. *Acta Biotheoretica* 58, 217–232. doi: 10.1007/s10441-010-9097-6
- Chelliah, V., Juty, N., Ajmera, I., Ali, R., Dumousseau, M., Glont, M., et al. (2015). Biomodels: ten-year anniversary. *Nucl. Acids Res.* 43(Database issue):D542–D548. doi: 10.1093/nar/gku1181
- Cimatti, A., Clarke, E. M., Giunchiglia, E., Giunchiglia, F., Pistore, M., Roveri, M., et al. (2002). “NuSMV2: an OpenSource tool for symbolic model checking,” in *Computer Aided Verification, Volume 2404 of Lecture Notes in Computer Science*, eds E. Brinksma and K. G. Larsen (Copenhagen: Springer), 359–364.
- Clarke, E. M., Faeder, J. R., Langmead, C. J., Harris, L. A., Jha, S. K., and Legay, A. (2008). “Statistical model checking in BioLab: applications to the automated analysis of T-cell receptor signaling pathway,” in *Computational Methods in Systems Biology, Volume 5307 of Lecture Notes in Bioinformatics*, eds M. Heiner and A. M. Uhrmacher (Rostock: Springer), 231–250. doi: 10.1007/978-3-540-88562-7_18
- Clarke, E. M., Grumberg, O., and Peled, D. A. (1999). *Model Checking*. Cambridge, MA: MIT Press.
- Cohen, D. P. A., Martignetti, L., Robine, S., Barillot, E., Zinoviyev, A., and Calzone, L. (2015). Mathematical modelling of molecular pathways enabling tumour cell invasion and migration. *PLoS Comput. Biol.* 11:e1004571. doi: 10.1371/journal.pcbi.1004571
- Conroy, B. D., Herek, T. A., Shew, T. D., Latner, M., Larson, J. J., Allen, L., et al. (2014). Design, assessment, and *in vivo* evaluation of a computational model illustrating the role of CAV1 in CD4(+) T-lymphocytes. *Front. Immunol.* 5:599. doi: 10.3389/fimmu.2014.00599
- Corblin, F., Fanchon, E., and Trilling, L. (2010). Applications of a formal approach to decipher discrete genetic networks. *BMC Bioinformatics* 11:385. doi: 10.1186/1471-2105-11-385
- Crespo, I., Krishna, A., Le Béche, A., and del Sol, A. (2013). Predicting missing expression values in gene regulatory networks using a discrete logic modeling optimization guided by network stable states. *Nucl. Acids Res.* 41:e8. doi: 10.1093/nar/gks785
- Davidich, M., and Bornholdt, S. (2008). The transition from differential equations to Boolean networks: a case study in simplifying a regulatory network model. *J. Theor. Biol.* 255, 269–277. doi: 10.1016/j.jtbi.2008.07.020
- Dubrova, E., and Teslenko, M. (2011). A SAT-based algorithm for finding attractors in synchronous boolean networks. *IEEE/ACM Trans. Comput. Biol. Bioinform.* 8, 1393–1399. doi: 10.1109/tcbb.2010.20
- Fauré, A., Naldi, A., Chaouiya, C., and Thieffry, D. (2006). Dynamical analysis of a generic boolean model for the control of the mammalian cell cycle. *Bioinformatics* 22, e124–e131. doi: 10.1093/bioinformatics/btl210
- Fauré, A., Naldi, A., Lopez, F., Chaouiya, C., Ciliberto, A., and Thieffry, D. (2009). Modular logical modelling of the budding yeast cell cycle. *Mol. Biosyst.* 5, 1787–1796. doi: 10.1039/b910101m
- Fauré, A., and Thieffry, D. (2009). Logical modelling of cell cycle control in eukaryotes: a comparative study. *Mol. Biosyst.* 5, 1569–1581. doi: 10.1039/b907562n
- Fauré, A., Vreede, B. M. I., Sucena, E., and Chaouiya, C. (2014). A discrete model of drosophila eggshell patterning reveals cell-autonomous and juxtacrine effects. *PLoS Comput. Biol.* 10:e1003527. doi: 10.1371/journal.pcbi.1003527
- Ferrell, J. E. Jr., Tsai, T. Y.-C., and Yang, Q. (2011). Modeling the cell cycle: why do certain circuits oscillate? *Cell* 144, 874–885. doi: 10.1016/j.cell.2011.03.006
- Flobak, A., Baudot, A., Remy, E., Thommesen, L., Thieffry, D., Kuiper, M., et al. (2015). Discovery of drug synergies in gastric cancer cells predicted by logical modeling. *PLoS Comput. Biol.* 11:e1004426. doi: 10.1371/journal.pcbi.1004426
- Garg, A., Di Cara, A., Xenarios, I., Mendoza, L., and De Micheli, G. (2008). Synchronous versus asynchronous modeling of gene regulatory networks. *Bioinformatics* 24, 1917–1925. doi: 10.1093/bioinformatics/btn336
- Garg, A., Mohanram, K., Di Cara, A., De Micheli, G., and Xenarios, I. (2009). Modeling stochasticity and robustness in gene regulatory networks. *Bioinformatics* 25, i101–i109. doi: 10.1093/bioinformatics/btp214
- Glass, L., and Kauffman, S. A. (1973). The logical analysis of continuous, non-linear biochemical control networks. *J. Theor. Biol.* 39, 103–129. doi: 10.1016/0022-5193(73)90208-7
- González, A., Chaouiya, C., and Thieffry, D. (2008). Logical modelling of the role of the hh pathway in the patterning of the drosophila wing disc. *Bioinformatics* 24, i234–i240. doi: 10.1093/bioinformatics/btn266
- Grieco, L., Calzone, L., Bernard-Pierrot, I., Radvanyi, F., Kahn-Perlès, B., and Thieffry, D. (2013). Integrative modelling of the influence of MAPK network on cancer cell fate decision. *PLoS Comput. Biol.* 9:e1003286. doi: 10.1371/journal.pcbi.1003286
- Helikar, T., Kochi, N., Kowal, B., Dimri, M., Naramura, M., Raja, S. M., et al. (2013a). A comprehensive, multi-scale dynamical model of ErbB receptor signal transduction in human mammary epithelial cells. *PLoS ONE* 8:e61757. doi: 10.1371/journal.pone.0061757
- Helikar, T., Kowal, B., McClenathan, S., Bruckner, M., Rowley, T., Madrahimov, A., et al. (2012). The Cell Collective: toward an open and collaborative approach to systems biology. *BMC Syst. Biol.* 6:96. doi: 10.1186/1752-0509-6-96
- Helikar, T., Kowal, B., and Rogers, J. A. (2013b). A cell simulator platform: the cell collective. *Clin. Pharmacol. Ther.* 93, 393–395. doi: 10.1038/clpt.2013.41
- Helikar, T., and Rogers, J. A. (2009). Chemchains: a platform for simulation and analysis of biochemical networks aimed to laboratory scientists. *BMC Syst. Biol.* 3:58. doi: 10.1186/1752-0509-3-58
- Hinton, A., Kwiatkowska, M., Norman, G., and Parker, D. (2006). “PRISM: a tool for automatic verification of probabilistic systems,” in *Tools and Algorithms for the Construction and Analysis of Systems, Volume 3920 of Lecture Notes in Computer Science*, eds H. Hermanns and J. Palsberg (Vienna: Springer), 441–444. doi: 10.1007/11691372_29
- Huang, S., Ernberg, I., and Kauffman, S. (2009). Cancer attractors: a systems view of tumors from a gene network dynamics and developmental perspective. *Semin. Cell Dev. Biol.* 20, 869–876. doi: 10.1016/j.semcdb.2009.07.003
- Hucka, M., Finney, A., Sauro, H. M., Bolouri, H., Doyle, J. C., Kitano, H., et al. (2003). The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models. *Bioinformatics* 19, 524–531. doi: 10.1093/bioinformatics/btg015
- Irons, D. J. (2009). Logical analysis of the budding yeast cell cycle. *J. Theor. Biol.* 257, 543–559. doi: 10.1016/j.jtbi.2008.12.028
- Jacob, F., and Monod, J. (1961). Genetic regulatory mechanisms in the synthesis of proteins. *J. Mol. Biol.* 3, 318–356. doi: 10.1016/s0022-2836(61)80072-7
- Kauffman, S. A. (1969). Metabolic stability and epigenesis in randomly constructed genetic nets. *J. Theor. Biol.* 22, 437–467. doi: 10.1016/0022-5193(69)90015-0
- Kauffman, S. A. (1993). *The Origins of Order: Self-Organization and Selection in Evolution*. New York, NY: Oxford University Press.
- Keller, R., Klein, M., Thomas, M., Dräger, A., Metzger, U., Templin, M. F., et al. (2016). Coordinating role of RXR α in downregulating hepatic detoxification during inflammation revealed by fuzzy-logic modeling. *PLoS Comput. Biol.* 12:e1004431. doi: 10.1371/journal.pcbi.1004431
- Kirouac, D. C., Du, J. Y., Lahdenranta, J., Overland, R., Yarar, D., Paragas, V., et al. (2013). Computational modeling of ERBB2-amplified breast cancer identifies combined ErbB2/3 blockade as superior to the combination of MEK and AKT inhibitors. *Sci. Signal.* 6:ra68. doi: 10.1126/scisignal.2004008
- Klamt, S., Haus, U.-U., and Theis, F. (2009). Hypergraphs and cellular networks. *PLoS Comput. Biol.* 5:e1000385. doi: 10.1371/journal.pcbi.1000385
- Klarner, H., Bockmayr, A., and Siebert, H. (2015). Computing maximal and minimal trap spaces of Boolean networks. *Nat. Comput.* 14, 535–544. doi: 10.1007/s11047-015-9520-7
- Le Novère, N. (2015). Quantitative and logic modelling of molecular and gene networks. *Nat. Rev. Genet.* 16, 146–158. doi: 10.1038/nrg3885
- Le Novère, N., Finney, A., Hucka, M., Bhalla, U. S., Campagne, F., Collado-Vides, J., et al. (2005). Minimum information requested in the annotation of biochemical models (MIRIAM). *Nat. Biotechnol.* 23, 1509–1515.
- Li, F., Long, T., Lu, Y., Ouyang, Q., and Tang, C. (2004). The yeast cell-cycle network is robustly designed. *Proc. Natl. Acad. Sci. U.S.A.* 101, 4781–4786. doi: 10.1073/pnas.0305937101
- Lomuscio, A., Pecheur, C., and Raimondi, F. (2007). “Automatic verification of knowledge and time with NuSMV,” in *International Joint Conference on Artificial Intelligence*, ed M. M. Veloso (Hyderabad: AAAI), 1384–1389.
- MacNamara, A., Terfve, C., Henriques, D., Bernabé, B. P., and Saez-Rodriguez, J. (2012). State-time spectrum of signal transduction logic models. *Phys. Biol.* 9:045003. doi: 10.1088/1478-3975/9/4/045003

- Madrahimov, A., Helikar, T., Kowal, B., Lu, G., and Rogers, J. (2013). Dynamics of influenza virus and human host interactions during infection and replication cycle. *Bull. Math. Biol.* 75, 988–1011. doi: 10.1007/s11538-012-9777-2
- Martinez-Sanchez, M. E., Mendoza, L., Villarreal, C., and Álvarez-Buylla, E. R. (2015). A minimal regulatory network of extrinsic and intrinsic factors recovers observed patterns of CD4+ T cell differentiation and plasticity. *PLoS Comput. Biol.* 11:e1004324. doi: 10.1371/journal.pcbi.1004324
- Martínez-Sosa, P., and Mendoza, L. (2013). The regulatory network that controls the differentiation of T lymphocytes. *BioSystems* 113, 96–103. doi: 10.1016/j.biosystems.2013.05.007
- Mendes, N. D., Monteiro, P. T., Carneiro, J., Remy, E., and Chaouiya, C. (2014). Quantification of reachable attractors in asynchronous discrete dynamics. arXiv:1411.3539 [cs.DM].
- Mendoza, L. (2006). A network model for the control of the differentiation process in Th cells. *BioSystems* 84, 101–114. doi: 10.1016/j.biosystems.2005.10.004
- Mendoza, L., and Xenarios, I. (2006). A method for the generation of standardized qualitative dynamical systems of regulatory networks. *Theor. Biol. Med. Model.* 3:13. doi: 10.1186/1742-4682-3-13
- Miskov-Zivanov, N., Turner, M. S., Kane, L. P., Morel, P. A., and Faeder, J. R. (2013). The duration of T cell stimulation is a critical determinant of cell fate and plasticity. *Sci. Signal.* 6:ra97. doi: 10.1126/scisignal.2004217
- Mombach, J. C. M., Bugs, C. A., and Chaouiya, C. (2014). Modelling the onset of senescence at the G1/S cell cycle checkpoint. *BMC Genomics* 15(Suppl. 7):S7. doi: 10.1186/1471-2164-15-S7-S7
- Monod, J., and Jacob, F. (1961). Teleonomic mechanisms in cellular metabolism, growth, and differentiation. *Cold Spring Harb. Symp. Quan. Biol.* 26, 389–401. doi: 10.1101/SQB.1961.026.01.048
- Monteiro, P. T., and Chaouiya, C. (2012). “Efficient verification for logical models of regulatory networks,” in *PACBB, Volume 154 of Advances in Intelligent and Soft Computing*, eds M. P. Rocha, N. Luscombe, F. Fdez-Riverola, J. M. Corchado Rodríguez (Berlin; Heidelberg: Springer), 259–267.
- Morris, M. K., Saez-Rodriguez, J., Clarke, D. C., Sorger, P. K., and Lauffenburger, D. A. (2011). Training signaling pathway maps to biochemical data with constrained fuzzy logic: quantitative analysis of liver cell responses to inflammatory stimuli. *PLoS Comput. Biol.* 7:e1001099. doi: 10.1371/journal.pcbi.1001099
- Morris, M. K., Saez-Rodriguez, J., Sorger, P. K., and Lauffenburger, D. A. (2010). Logic-based models for the analysis of cell signaling networks. *Biochemistry* 49, 3216–3224. doi: 10.1021/bi902202q
- Murphy, K., Travers, P., Walport, M., and Janeway, C. (2012). *Janeway's Immunology*. New York, NY: Garland Science.
- Nakayama, S., Takahashi, H., Kanno, Y., and O'Shea, J. J. (2012). Helper T cell diversity and plasticity. *Curr. Opin. Immunol.* 24, 297–302. doi: 10.1016/j.coi.2012.01.014
- Naldi, A., Carneiro, J., Chaouiya, C., and Thieffry, D. (2010). Diversity and plasticity of Th cell types predicted from regulatory network modelling. *PLoS Comput. Biol.* 6:e1000912. doi: 10.1371/journal.pcbi.1000912
- Naldi, A., Monteiro, P. T., and Chaouiya, C. (2012). “Efficient handling of large signalling-regulatory networks by focusing on their core control,” in *Computational Methods in Systems Biology, Volume 7605 of Lecture Notes in Computer Science*, eds D. Gilbert and M. Heiner (London: Springer), 288–306. doi: 10.1007/978-3-642-33636-2_17
- Naldi, A., Monteiro, P. T., Müssel, C., Consortium for Logical Models and Tools, Kestler, H. A., Thieffry, D., et al. (2015). Cooperative development of logical modelling standards and tools with CoLoMoTo. *Bioinformatics* 31, 1154–1159. doi: 10.1093/bioinformatics/btv013
- Naldi, A., Remy, E., Thieffry, D., and Chaouiya, C. (2011). Dynamically consistent reduction of logical regulatory graphs. *Theor. Comput. Sci.* 412, 2207–2218. doi: 10.1016/j.tcs.2010.10.021
- Naldi, A., Thieffry, D., and Chaouiya, C. (2007). “Decision diagrams for the representation and analysis of logical models of genetic networks,” in *Computational Methods in Systems Biology, Volume 4695 of Lecture Notes in Computer Science*, eds M. Calder and S. Gilmore (Edinburgh: Springer), 233–247. doi: 10.1007/978-3-540-75140-3_16
- Novák, B., and Tyson, J. J. (2004). A model for restriction point control of the mammalian cell cycle. *J. Theor. Biol.* 230, 563–579. doi: 10.1016/j.jtbi.2004.04.039
- Oyeyemi, O. J., Davies, O., Robertson, D. L., and Schwartz, J.-M. (2015). A logical model of HIV-1 interactions with the T-cell activation signalling pathway. *Bioinformatics* 31, 1075–1083. doi: 10.1093/bioinformatics/btu787
- Puniya, B. L., Allen, L., Hochfelder, C., Majumder, M., and Helikar, T. (2016). Systems perturbation analysis of a large scale signal transduction model reveals potentially influential candidates for cancer therapeutics. *Front. Bioeng. Biotechnol.* 4:10. doi: 10.3389/fbioe.2016.00010
- Remy, E., Mossé, B., Chaouiya, C., and Thieffry, D. (2003). A description of dynamical graphs associated to elementary regulatory circuits. *Bioinformatics* 19(Suppl. 2), ii172–ii178. doi: 10.1093/bioinformatics/btg1075
- Remy, E., Rebouissou, S., Chaouiya, C., Zinoviyev, A., Radvanyi, F., and Calzone, L. (2015). A modeling approach to explain mutually exclusive and co-occurring genetic alterations in bladder tumorigenesis. *Cancer Res.* 75, 4042–4052. doi: 10.1158/0008-5472.CAN-15-0602
- Saadatpour, A., Albert, R., and Reluga, T. C. (2013). A reduction method for Boolean network models proven to conserve attractors. *SIAM J. Appl. Dyn. Syst.* 12, 1997–2011. doi: 10.1137/13090537x
- Saadatpour, A., Wang, R.-S., Liao, A., Liu, X., Loughran, T. P., Albert, I., et al. (2011). Dynamical and structural analysis of a T cell survival network identifies novel candidate therapeutic targets for large granular lymphocyte leukemia. *PLoS Comput. Biol.* 7:e1002267. doi: 10.1371/journal.pcbi.1002267
- Saez-Rodriguez, J., Alexopoulos, L. G., Epperlein, J., Samaga, R., Lauffenburger, D. A., Klamt, S., et al. (2009). Discrete logic modelling as a means to link protein signalling networks with functional analysis of mammalian signal transduction. *Mol. Syst. Biol.* 5:331. doi: 10.1038/msb.2009.87
- Saez-Rodriguez, J., Simeoni, L., Lindquist, J. A., Hemenway, R., Bommhardt, U., Arndt, B., et al. (2007). A logical model provides insights into T cell receptor signaling. *PLoS Comput. Biol.* 3:e163. doi: 10.1371/journal.pcbi.0030163
- Sahin, O., Fröhlich, H., Löbke, C., Korf, U., Burmester, S., Majety, M., et al. (2009). Modeling ERBB receptor-regulated G1/S transition to find novel targets for de novo trastuzumab resistance. *BMC Syst. Biol.* 3:1. doi: 10.1186/1752-0509-3-1
- Samaga, R., and Klamt, S. (2013). Modeling approaches for qualitative and semi-quantitative analysis of cellular signaling networks. *Cell Commun. Signal* 11:43. doi: 10.1186/1478-811X-11-43
- Sánchez, L., Chaouiya, C., and Thieffry, D. (2008). Segmenting the fly embryo: logical analysis of the role of the segment polarity cross-regulatory module. *Int. J. Dev. Biol.* 52, 1059–1075. doi: 10.1387/ijdb.072439ls
- Schlatter, R., Schlich, K., Avalos Vizcarra, I., Scheurich, P., Sauter, T., Borner, C., et al. (2009). ON/OFF and beyond—a Boolean model of apoptosis. *PLoS Comput. Biol.* 5:e1000595. doi: 10.1371/journal.pcbi.1000595
- Shmulevich, I., Dougherty, E. R., Kim, S., and Zhang, W. (2002). Probabilistic boolean networks: a rule-based uncertainty model for gene regulatory networks. *Bioinformatics* 18, 261–274. doi: 10.1093/bioinformatics/18.2.261
- Stoll, G., Viara, E., Barillot, E., and Calzone, L. (2012). Continuous time Boolean modeling for biological signaling: application of Gillespie algorithm. *BMC Syst. Biol.* 6:116. doi: 10.1186/1752-0509-6-116
- Sugita, M. (1963). Functional analysis of chemical systems *in vivo* using a logical circuit equivalent. II. The idea of a molecular automation. *J. Theor. Biol.* 4, 179–192. doi: 10.1016/0022-5193(63)90087-0
- Terfve, C., Cokelaer, T., Henriques, D., MacNamara, A., Gonçalves, E., Morris, M. K., et al. (2012). CellNOptR: a flexible toolkit to train protein signaling networks to data using multiple logic formalisms. *BMC Syst. Biol.* 6:133. doi: 10.1186/1752-0509-6-133
- Terfve, C. D. A., Wilkes, E. H., Casado, P., Cutillas, P. R., and Saez-Rodriguez, J. (2015). Large-scale models of signal propagation in human cells derived from discovery phosphoproteomic data. *Nat. Commun.* 6:8033. doi: 10.1038/ncomms9033
- Thieffry, D. (2007). Dynamical roles of biological regulatory circuits. *Brief. Bioinform.* 8, 220–225. doi: 10.1093/bib/bbm028
- Thieffry, D., and Thomas, R. (1995). Dynamical behaviour of biological regulatory networks—II. Immunity control in bacteriophage lambda. *Bull. Math. Biol.* 57, 277–297.
- Thomas, R. (1973). Boolean formalization of genetic control circuits. *J. Theor. Biol.* 42, 563–585. doi: 10.1016/0022-5193(73)90247-6
- Thomas, R. (1978). Logical analysis of systems comprising feedback loops. *J. Theor. Biol.* 73, 631–656. doi: 10.1016/0022-5193(78)90127-3

- Thomas, R., and D'Ari, R. (1990). *Biological Feedback*. Boca Raton, FL: CRC Press.
- Thomas, R., Gathoye, A. M., and Lambert, L. (1976). A complex control circuit regulation of immunity in temperate bacteriophages. *Eur. J. Biochem.* 71, 211–227. doi: 10.1111/j.1432-1033.1976.tb11108.x
- Todd, R. G., and Helikar, T. (2012). Ergodic sets as cell phenotype of budding yeast cell cycle. *PLoS ONE* 7:e45780. doi: 10.1371/journal.pone.0045780
- Traynard, P., Fauré, A., Fages, F., and Thieffry, D. (2015). Logical modeling of the mammalian cell cycle. *viXra*:1512.0337.
- Tyson, J. J., and Novák, B. (2015). Models in biology: lessons from modeling regulation of the eukaryotic cell cycle. *BMC Biol.* 13:46. doi: 10.1186/s12915-015-0158-9
- Videla, S., Guziolowski, C., Eduati, F., Thiele, S., Gebser, M., Nicolas, J., et al. (2015). Learning Boolean logic models of signaling networks with ASP. *Theor. Comp. Sci.* 599, 79–101. doi: 10.1016/j.tcs.2014.06.022
- Waltemath, D., Adams, R., Bergmann, F. T., Hucka, M., Kolpakov, F., Miller, A. K., et al. (2011). Reproducible computational biology experiments with SED-ML—the simulation experiment description markup language. *BMC Syst. Biol.* 5:198. doi: 10.1186/1752-0509-5-198
- Wittmann, D. M., Krumsiek, J., Saez-Rodriguez, J., Lauffenburger, D. A., Klamt, S., and Theis, F. J. (2009). Transforming Boolean models to continuous models: methodology and application to t-cell receptor signaling. *BMC Syst. Biol.* 3:98. doi: 10.1186/1752-0509-3-98
- Zañudo, J. G. T., and Albert, R. (2013). An effective network reduction approach to find the dynamical repertoire of discrete dynamic networks. *Chaos* 23:025111. doi: 10.1063/1.4809777
- Zhang, R., Shah, M. V., Yang, J., Nyland, S. B., Liu, X., Yun, J. K., et al. (2008). Network model of survival signaling in large granular lymphocyte leukemia. *Proc. Natl. Acad. Sci. U.S.A.* 105, 16308–16313. doi: 10.1073/pnas.0806447105

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2016 Abou-Jaoudé, Traynard, Monteiro, Saez-Rodriguez, Helikar, Thieffry and Chaouiya. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



Elementary Vectors and Conformal Sums in Polyhedral Geometry and their Relevance for Metabolic Pathway Analysis

Stefan Müller* and Georg Regensburger

Radon Institute for Computational and Applied Mathematics, Austrian Academy of Sciences, Linz, Austria

OPEN ACCESS

Edited by:

Alberto Marin-Sanguino,
Technical University of Munich,
Germany

Reviewed by:

Jesus Picó,
Universitat Politècnica de València,
Spain
Ruriko Yoshida,
University of Kentucky, USA
Carsten Wiuf,
University of Copenhagen, Denmark

*Correspondence:

Stefan Müller
stefan.mueller@ricam.oeaw.ac.at

Specialty section:

This article was submitted to
Systems Biology,
a section of the journal
Frontiers in Genetics

Received: 29 November 2015

Accepted: 02 May 2016

Published: 24 May 2016

Citation:

Müller S and Regensburger G (2016)
Elementary Vectors and Conformal
Sums in Polyhedral Geometry and
their Relevance for Metabolic Pathway
Analysis. *Front. Genet.* 7:90.
doi: 10.3389/fgene.2016.00090

A fundamental result in metabolic pathway analysis states that every flux mode can be decomposed into a sum of elementary modes. However, only a decomposition without cancelations is biochemically meaningful, since a reversible reaction cannot have different directions in the contributing elementary modes. This essential requirement has been largely overlooked by the metabolic pathway community. Indeed, every flux mode can be decomposed into elementary modes without cancelations. The result is an immediate consequence of a theorem by Rockafellar which states that every element of a linear subspace is a conformal sum (a sum without cancelations) of elementary vectors (support-minimal vectors). In this work, we extend the theorem, first to “subspace cones” and then to general polyhedral cones and polyhedra. Thereby, we refine Minkowski’s and Carathéodory’s theorems, two fundamental results in polyhedral geometry. We note that, in general, elementary vectors need not be support-minimal; in fact, they are conformally non-decomposable and form a unique minimal set of conformal generators. Our treatment is mathematically rigorous, but suitable for systems biologists, since we give self-contained proofs for our results and use concepts motivated by metabolic pathway analysis. In particular, we study cones defined by linear subspaces and nonnegativity conditions — like the flux cone — and use them to analyze general polyhedral cones and polyhedra. Finally, we review applications of elementary vectors and conformal sums in metabolic pathway analysis.

Keywords: Minkowski’s theorem, Carathéodory’s theorem, s-cone, polyhedral cone, polyhedron, conformal generators

1. INTRODUCTION

Cellular metabolism is the set of biochemical reactions which transform nutrients from the environment into all the biomolecules a living cell consists of. Most metabolic reactions are catalyzed by enzymes, the expression and activity of which is controlled by gene and allosteric regulation, respectively.

A metabolic network together with enzymatic reaction rates gives rise to a nonlinear dynamical system for the metabolite concentrations. However, for genome-scale networks, quantitative knowledge of the underlying kinetics is not available, and a mathematical analysis is not practicable. Instead, one considers only stoichiometric information and studies the system of linear equalities

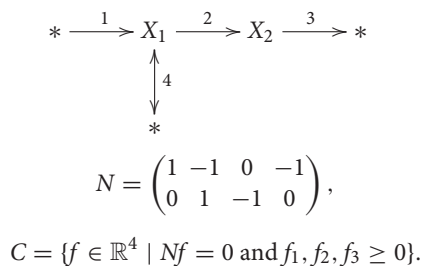
and inequalities for the fluxes (net reaction rates), arising from the pseudo steady-state assumption and irreversibility constraints.

A metabolic network is given by n internal metabolites, r reactions, and the corresponding stoichiometric matrix $N \in \mathbb{R}^{n \times r}$, which contains the net stoichiometric coefficients of each metabolite in each reaction. The set of irreversible reactions is given by $\mathcal{I} \subseteq \{1, \dots, r\}$. One is interested in the *flux cone*

$$C = \{f \in \mathbb{R}^r \mid Nf = 0 \text{ and } f_i \geq 0 \text{ for } i \in \mathcal{I}\},$$

which is a polyhedral cone defined by the null-space of the stoichiometric matrix and nonnegativity conditions. Its elements are called *flux modes*.

As a running example, we consider a small network, taken from Schuster et al. (2002), the corresponding stoichiometric matrix, and the resulting flux cone:



The network consists of two internal metabolites X_1, X_2 and four chemical reactions. Reaction 1 imports X_1 from the environment (indicated by the symbol $*$) which yields the first column $(1, 0)^T$ of the stoichiometric matrix N . Reaction 2 transforms X_1 into X_2 which gives the column $(-1, 1)^T$, and reaction 3 exports X_2 which gives $(0, -1)^T$. The first three reactions are assumed to be irreversible which yields the nonnegativity constraints $f_1, f_2, f_3 \geq 0$ in the definition of the flux cone C . Finally, reaction 4 is reversible and exports/imports X_1 .

Metabolic pathway analysis aims to identify biochemically/biologically/biotechnologically meaningful routes in a network, in particular, the smallest routes. Several definitions for minimal metabolic pathways have been given in the literature, with *elementary modes* (EMs) being the fundamental concept both biologically and mathematically Klamt and Stelling (2003); Llaneras and Picó (2010). Formally, EMs are defined as support-minimal (or, equivalently, support-wise non-decomposable) flux modes Schuster and Hilgetag (1994); Schuster et al. (2002). Clearly, a positive multiple of an EM is also an EM since it fulfills the steady-state condition and the irreversibility constraints.

In the example, the EMs are given by $e^1 = (1, 0, 0, 1)^T$, $e^2 = (0, 1, 1, -1)^T$, $e^3 = (1, 1, 1, 0)^T$, and their positive multiples. It is easy to check that e^1, e^2 , and e^3 are flux modes (elements of the flux cone) and support-minimal. Note that $e^3 = e^1 + e^2$.

A fundamental result in metabolic pathway analysis states that every flux mode can be decomposed into a sum of EMs Schuster et al. (2002). However, only a decomposition without cancelations is biochemically meaningful, since a reversible reaction cannot have different directions in the

contributing EMs. This essential requirement has been largely overlooked by the metabolic pathway community. Indeed, as we will show in this work, every flux mode can be decomposed into EMs *without cancelations*, that is,

- (0) if a component of the flux mode is zero, then this component is zero in the contributing EMs,
- (+) if a component of the flux mode is positive, then this component is positive or zero in the contributing EMs,
- (−) if a component of the flux mode is negative, then this component is negative or zero in the contributing EMs.

In mathematical terms, every nonzero element of a “subspace cone” (defined by a linear subspace and nonnegativity conditions) is a conformal sum of elementary vectors, cf. Theorem 3. The result is stated in Urbanczik and Wagner (2005) and Urbanczik (2007); part (0) has been shown in Schuster et al. (2002) and guarantees a decomposition without cancelations in a weaker sense Llaneras and Picó (2010); Zanghellini et al. (2013).

In the example, the flux mode $f = (2, 1, 1, 1)^T$ can be decomposed into EMs in two ways:

$$\begin{aligned} f &= \begin{pmatrix} 2 \\ 1 \\ 1 \\ 1 \end{pmatrix} = 2e^1 + e^2 = \begin{pmatrix} 2 \\ 0 \\ 0 \\ 2 \end{pmatrix} + \begin{pmatrix} 0 \\ 1 \\ 1 \\ -1 \end{pmatrix} \\ &= e^1 + e^3 = \begin{pmatrix} 1 \\ 0 \\ 0 \\ 1 \end{pmatrix} + \begin{pmatrix} 1 \\ 1 \\ 1 \\ 0 \end{pmatrix}. \end{aligned}$$

The first sum involves a cancelation in the last component of the flux. The last reaction is reversible, however, it cannot have a net rate in different directions at the same time. Hence, only the second sum is biochemically meaningful. As stated above, a decomposition without cancelations is always possible.

In convex analysis, elementary vectors of a linear subspace were introduced as support-minimal vectors by Rockafellar in 1969. He proves that every vector is a conformal sum (originally called harmonious superposition) of elementary vectors (Rockafellar, 1969, Theorem 1). For proofs and generalizations in the settings of polyhedral geometry and oriented matroids (see Ziegler, 1995, Lemma 6.7) and (Bachem and Kern, 1992, Theorem 5.36). Rockafellar points out that this result is easily shown to be equivalent to Minkowski’s theorem Minkowski (1896) for pointed polyhedral cones, stating that every nonzero vector is a nonnegative linear combination of extreme vectors. Moreover, the result immediately implies Carathéodory’s theorem Carathéodory (1911), stating that the number of extreme vectors in such a nonnegative linear combination need not exceed the dimension of the cone. In fact, Rockafellar writes: “This is even a convenient route for attaining various important facts about polyhedral convex cones, since the direct proof [...] for Theorem 1 is so elementary.”

In metabolic pathway analysis, decompositions without cancelations were introduced by Urbanczik and Wagner (2005). The corresponding elementary vectors are defined by intersecting a polyhedral cone with all closed orthants of maximal dimension.

By applying Minkowski's theorem for pointed polyhedral cones, every vector is a sum of extreme vectors without cancelations. Urbanczik further extended this approach to polyhedra arising from flux cones and inhomogeneous constraints Urbanczik (2007).

In polyhedral geometry, it seems that conformal decompositions of general cones and polyhedra have not yet been studied. In this work, following Rockafellar, we first extend his result to cones defined by linear subspaces and nonnegativity conditions (Theorem 3). For subspace cones, support-minimality is equivalent to conformal non-decomposability. As it turns out, for general polyhedral cones, elementary vectors have to be defined as conformally non-decomposable vectors. However, these are in one-to-one correspondence with elementary vectors of a higher-dimensional subspace cone, and, by our result for subspace cones, we obtain a conformal refinement of Minkowski's and Carathéodory's theorems for polyhedral cones (Theorem 8). In particular, there is an upper bound on the number of elementary vectors needed in a conformal decomposition of a vector. Finally, by taking into account vertices and conformal convex combinations, we further extend our result to polyhedra (Theorem 13). We note that elementary vectors do not form a minimal generating set (of an s -cone, a general polyhedral cone, or a polyhedron). However, they form a unique minimal set of *conformal* generators (Proposition 17).

2. DEFINITIONS

We denote the nonnegative real numbers by \mathbb{R}_{\geq} . For $x \in \mathbb{R}^n$, we write $x \geq 0$ if $x \in \mathbb{R}_{\geq}^n$. Further, we denote the *support* of a vector $x \in \mathbb{R}^n$ by $\text{supp}(x) = \{i \mid x_i \neq 0\}$.

2.1. Sign Vectors

For $x \in \mathbb{R}^n$, we define the *sign vector* $\text{sign}(x) \in \{-, 0, +\}^n$ by applying the sign function component-wise, that is, $\text{sign}(x)_i = \text{sign}(x_i)$ for $i = 1, \dots, n$. The relations $0 < -$ and $0 < +$ induce a partial order on $\{-, 0, +\}^n$: for $X, Y \in \{-, 0, +\}^n$, we write $X \leq Y$ if the inequality holds component-wise. For $x, y \in \mathbb{R}^n$, we say that x *conforms* to y , if $\text{sign}(x) \leq \text{sign}(y)$. For example, let $x = (-1, 0, 2)^T$ and $y = (-2, -1, 1)$. Then,

$$\text{sign} \begin{pmatrix} -1 \\ 0 \\ 2 \end{pmatrix} = \begin{pmatrix} - \\ 0 \\ + \end{pmatrix} \leq \begin{pmatrix} - \\ - \\ + \end{pmatrix} = \text{sign} \begin{pmatrix} -2 \\ -1 \\ 1 \end{pmatrix},$$

that is, $\text{sign}(x) \leq \text{sign}(y)$, and x conforms to y . Let $X \in \{-, 0, +\}^n$. The corresponding closed orthant $O \subset \mathbb{R}^n$ is defined as $O = \{x \mid \text{sign}(x) \leq X\}$.

2.2. Convex Cones

A nonempty subset C of a vector space is a *convex cone*, if

$$x, y \in C \text{ and } \mu, \nu > 0 \text{ imply } \mu x + \nu y \in C,$$

or, equivalently, if

$$\lambda C = C \text{ for all } \lambda > 0 \text{ and } C + C = C.$$

A convex cone C is called *pointed* if $C \cap -C = \{0\}$. It is *polyhedral* if

$$C = \{x \mid Ax \geq 0\} \quad \text{for some } A \in \mathbb{R}^{m \times r},$$

that is, if it is defined by finitely many homogeneous inequalities. Hence, a polyhedral cone is pointed if and only if $\ker(A) = \{0\}$.

2.3. Special Vectors

We recall the definitions of support-minimal vectors and extreme vectors, which play an important role in both polyhedral geometry and metabolic pathway analysis. We also introduce support-wise non-decomposable vectors, which serve as elementary modes for flux cones (in the original definition), and conformally non-decomposable vectors, which serve as elementary vectors for general polyhedral cones (see Subsection 3.2).

Let C be a convex cone. A nonzero vector $x \in C$ is called

- *support-minimal*, if

for all nonzero $x' \in C$,

$$\text{supp}(x') \subseteq \text{supp}(x) \text{ implies } \text{supp}(x') = \text{supp}(x), \quad (\text{SM})$$

- *support-wise non-decomposable*, if

for all nonzero $x^1, x^2 \in C$ with $\text{supp}(x^1), \text{supp}(x^2) \subseteq \text{supp}(x)$,

$$x = x^1 + x^2 \text{ implies } \text{supp}(x^1) = \text{supp}(x^2), \quad (\text{swND})$$

- *conformally non-decomposable*, if

for all nonzero $x^1, x^2 \in C$ with $\text{sign}(x^1), \text{sign}(x^2) \leq \text{sign}(x)$,

$$x = x^1 + x^2 \text{ implies } x^1 = \lambda x^2 \text{ with } \lambda > 0, \quad (\text{cND})$$

- and *extreme*, if

for all nonzero $x^1, x^2 \in C$,

$$x = x^1 + x^2 \text{ implies } x^1 = \lambda x^2 \text{ with } \lambda > 0. \quad (\text{EX})$$

From the definitions, we have the implications

$$\text{SM} \Rightarrow \text{swND} \Leftarrow \text{EX} \Rightarrow \text{cND}.$$

If $x \in C$ is extreme, then $\{\lambda x \mid \lambda > 0\}$ is called an extreme ray of C . In fact, C has an extreme ray if and only if C is pointed. If C is contained in a closed orthant (and hence pointed), we have the equivalence $\text{cND} \Leftrightarrow \text{EX}$.

3. MATHEMATICAL RESULTS

We start by extending a result on conformal decompositions into elementary vectors from linear subspaces to special cases of polyhedral cones, including flux cones in metabolic pathway analysis.

3.1. Linear Subspaces and S-cones

We consider linear subspaces with optional nonnegativity constraints as special cases of polyhedral cones. Let $S \subseteq \mathbb{R}^r$ be a linear subspace and $0 \leq d \leq r$. We define the resulting s-cone (subspace cone, special cone) as

$$C(S, d) = \left\{ \begin{pmatrix} x \\ y \end{pmatrix} \in \mathbb{R}^{(r-d)+d} \mid \begin{pmatrix} x \\ y \end{pmatrix} \in S, y \geq 0 \right\}.$$

Clearly, $C(S, 0) = S$ and $C(S, r) = S \cap \mathbb{R}_{\geq}^r$.

Definition 1. Let $C(S, d)$ be an s-cone. A vector $e \in C(S, d)$ is called elementary if it is support-minimal.

For linear subspaces, the definition of elementary vectors (EVs) as SM vectors was given in Rockafellar (1969). For flux cones, where $S = \ker(N)$, the definition of elementary modes (EMs) as SM vectors was given in Schuster et al. (2002). Interestingly, the choice of the same adjective for the closely related concepts of elementary vectors and elementary modes was coincidental Schuster (2015).

In the proofs of Theorem 3 and Propositions 4 and 5, we use the following argument.

Lemma 2. Let $C(S, d)$ be an s-cone and $x, x' \in C(S, d)$ be nonzero vectors which are not proportional. If $\text{supp}(x') \subseteq \text{supp}(x)$, then there exists a nonzero vector

$$x'' = x - \lambda x' \in C(S, d) \quad \text{with } \lambda \in \mathbb{R}$$

such that

$$\text{sign}(x'') \leq \text{sign}(x) \quad \text{and} \quad \text{supp}(x'') \subset \text{supp}(x).$$

If $\text{sign}(x') \leq \text{sign}(x)$, then $\lambda > 0$ in x'' .

Proof. Clearly, $x'' = x - \lambda x'$ is nonzero for all $\lambda \in \mathbb{R}$. There exists a largest $\lambda > 0$ (in case $\text{sign}(-x') \leq \text{sign}(x)$ a smallest $\lambda < 0$) such that $\text{sign}(x'') \leq \text{sign}(x)$. For this λ , $x'' \in C(S, d)$ and $\text{supp}(x'') \subset \text{supp}(x)$. \square

For linear subspaces, the following fundamental result was proved in Rockafellar (1969 Theorem 1). We extend it to s-cones.

Theorem 3. Let $C(S, d)$ be an s-cone. Every nonzero vector $x \in C(S, d)$ is a conformal sum of EVs. That is, there exists a finite set $E \subseteq C(S, d)$ of EVs such that

$$x = \sum_{e \in E} e \quad \text{with } \text{sign}(e) \leq \text{sign}(x).$$

The set E can be chosen such that its elements are linearly independent, in particular, they can be ordered such that every $e \in E$ has a component which is nonzero in e , but zero in its predecessors (in the ordered set). Then, $|E| \leq \dim(S)$ and $|E| \leq |\text{supp}(x)|$.

Proof. We proceed by induction on the cardinality of $\text{supp}(x)$.

Either, x is SM (and $E = \{x\}$) or there exists a nonzero vector $x' \in C(S, d)$ with $\text{supp}(x') \subset \text{supp}(x)$, but not necessarily with $\text{sign}(x') \leq \text{sign}(x)$. However, by Lemma 2, there exists a nonzero vector $x'' \in C(S, d)$ with $\text{sign}(x'') \leq \text{sign}(x)$ and $\text{supp}(x'') \subset \text{supp}(x)$. By the induction hypothesis, there exists a SM vector e^* with $\text{sign}(e^*) \leq \text{sign}(x'')$ and hence $\text{sign}(e^*) \leq \text{sign}(x)$. By Lemma 2 again, there exists a nonzero vector

$$x^* = x - \lambda e^* \in C(S, d) \quad \text{with } \lambda > 0$$

such that $\text{sign}(x^*) \leq \text{sign}(x)$ and $\text{supp}(x^*) \subset \text{supp}(x)$. By the induction hypothesis, there exists a finite set E^* of SM vectors such that

$$x^* = \sum_{e \in E^*} e \quad \text{with } \text{sign}(e) \leq \text{sign}(x^*)$$

and hence $\text{sign}(e) \leq \text{sign}(x)$. We have constructed a finite set $E = E^* \cup \{\lambda e^*\}$ of SM vectors such that

$$x = x^* + \lambda e^* = \sum_{e \in E^*} e + \lambda e^* = \sum_{e \in E} e \quad \text{with } \text{sign}(e) \leq \text{sign}(x).$$

By the induction hypothesis, the set E^* can be chosen such that its elements are linearly independent and ordered such that every $e \in E^*$ has a component which is nonzero in e , but zero in all its predecessors. By construction, λe^* has a component which is nonzero, but zero in x^* and hence in all $e \in E^*$. Obviously, the elements of $E = E^* \cup \{\lambda e^*\}$ are linearly independent and can be ordered accordingly. \square

The statement about the support of the EVs was too strong in Rockafellar (1969, Theorem 1). It was claimed that every EV has a nonzero component which is zero in all other EVs.*

Theorem 3 is a conformal refinement of Minkowski's and Carathéodory's theorems for s-cones. In fact, it remains to show that there are finitely many EVs.

Proposition 4. Let $C(S, d)$ be an s-cone. If two SM vectors $x, x' \in C(S, d)$ have the same sign vector, $\text{sign}(x) = \text{sign}(x')$, then $x = \lambda x'$ with $\lambda > 0$. As a consequence, there are finitely many SM vectors up to positive scalar multiples.

Proof. Assume there are two SM vectors with the same sign vector which are not proportional. Then, by Lemma 2, there exists a vector with smaller support. \square

* For a counterexample, consider the subspace $S = \ker(1, -1, -1, 1) \subseteq \mathbb{R}^4$. Its nonnegative EVs are

$$e^1 = \begin{pmatrix} 1 \\ 1 \\ 0 \\ 0 \end{pmatrix}, e^2 = \begin{pmatrix} 1 \\ 0 \\ 1 \\ 0 \end{pmatrix}, e^3 = \begin{pmatrix} 0 \\ 1 \\ 0 \\ 1 \end{pmatrix}, e^4 = \begin{pmatrix} 0 \\ 0 \\ 1 \\ 1 \end{pmatrix},$$

and their positive multiples. Then $x = (1, 2, 3, 4)^T$ is not a conformal sum of EVs with the claimed property. (Every conformal decomposition of x consists of at least 3 EVs, and every set of 3 EVs contains 1 EV which does not have a nonzero component which is zero in the other EVs.)

We conclude by showing that, for s-cones, EVs can be equivalently defined as SM, swND, or cND vectors.

Proposition 5. *For an s-cone, support-minimality, support-wise non-decomposability, and conformal non-decomposability are equivalent. That is,*

$$\text{s-cone: } SM \Leftrightarrow \text{swND} \Leftrightarrow \text{cND}.$$

Proof. $SM \Rightarrow \text{swND}$: By definition.

$\text{swND} \Rightarrow \text{cND}$: Let $C(S, d)$ be an s-cone and assume that $x \in C(S, d)$ is conformally decomposable, that is, $x = x^1 + x^2$ with nonzero $x^1, x^2 \in C(S, d)$, $\text{sign}(x^1), \text{sign}(x^2) \leq \text{sign}(x)$, and x^1, x^2 being not proportional. By Lemma 2, there exists a nonzero $x' = x - \lambda x^1 \in C(S, d)$ such that $\text{supp}(x') \subset \text{supp}(x)$. Hence $\text{supp}(x') \neq \text{supp}(x^1)$, and $x = x' + \lambda x^1$ is support-wise decomposable.

$\text{cND} \Rightarrow \text{SM}$: Let $C(S, d)$ be an s-cone and assume that $x \in C(S, d)$ is not SM, that is, there exists a nonzero $x' \in C(S, d)$ with $\text{supp}(x') \subset \text{supp}(x)$. Then, there exists a largest $\lambda > 0$ such that $x^1 = \frac{1}{2}x + \lambda x'$ and $x^2 = \frac{1}{2}x - \lambda x'$ fulfill $\text{sign}(x^1), \text{sign}(x^2) \leq \text{sign}(x)$. For this λ , either $\text{supp}(x^1) \subset \text{supp}(x)$ or $\text{supp}(x^2) \subset \text{supp}(x)$; in any case, $x^1, x^2 \in C(S, d)$ and $\text{supp}(x^1) \neq \text{supp}(x^2)$. Hence, $x = x^1 + x^2$ is conformally decomposable. \square

If an s-cone is contained in a closed orthant, then further $\text{cND} \Leftrightarrow \text{EX}$, and all definitions of special vectors are equivalent.

3.2. General Polyhedral Cones

Let C be a polyhedral cone, that is,

$$C = \{x \in \mathbb{R}^r \mid Ax \geq 0\} \quad \text{for some } A \in \mathbb{R}^{m \times r}.$$

For s-cones, we defined elementary vectors (EVs) via support-minimality which, in this case, turned out to be equivalent to conformal non-decomposability. For general polyhedral cones, only the latter concept allows to extend Theorem 3.

Definition 6. *Let C be a polyhedral cone. A vector $e \in C$ is called elementary if it is conformally non-decomposable.*

In order to apply Theorem 3, we define an s-cone related to a polyhedral cone C . We introduce the subspace

$$\tilde{S} = \left\{ \begin{pmatrix} x \\ Ax \end{pmatrix} \in \mathbb{R}^{r+m} \mid x \in \text{span}(C) \right\}$$

with $\dim(\tilde{S}) = \dim(C)$ and the s-cone

$$\begin{aligned} \tilde{C} &= C(\tilde{S}, m) \\ &= \left\{ \begin{pmatrix} x \\ Ax \end{pmatrix} \in \mathbb{R}^{r+m} \mid x \in \text{span}(C) \text{ and } Ax \geq 0 \right\} \\ &= \left\{ \begin{pmatrix} x \\ Ax \end{pmatrix} \in \mathbb{R}^{r+m} \mid x \in C \right\}. \end{aligned}$$

Hence,

$$x \in C \Leftrightarrow \begin{pmatrix} x \\ Ax \end{pmatrix} \in \tilde{C}.$$

Moreover, the cND vectors of C and \tilde{C} are in one-to-one correspondence.

Lemma 7. *Let $C = \{x \mid Ax \geq 0\}$ be a polyhedral cone and $\tilde{C} = \left\{ \begin{pmatrix} x \\ Ax \end{pmatrix} \mid Ax \geq 0 \right\}$ the related s-cone. Then,*

$$x \in C \text{ is cND} \Leftrightarrow \begin{pmatrix} x \\ Ax \end{pmatrix} \in \tilde{C} \text{ is cND}.$$

Proof. First, we show the equivalence of the premises in the definitions of conformal non-decomposability for C and \tilde{C} . Indeed,

$$\begin{aligned} x = x^1 + x^2 \text{ with } x^1, x^2 \in C \\ \Leftrightarrow \\ \begin{pmatrix} x \\ Ax \end{pmatrix} = \begin{pmatrix} x^1 \\ Ax^1 \end{pmatrix} + \begin{pmatrix} x^2 \\ Ax^2 \end{pmatrix} \text{ with } \begin{pmatrix} x^1 \\ Ax^1 \end{pmatrix}, \begin{pmatrix} x^2 \\ Ax^2 \end{pmatrix} \in \tilde{C}. \end{aligned}$$

Assuming $x = x^1 + x^2$ with $x^1, x^2 \in C$ (and hence $Ax^1, Ax^2, Ax \geq 0$), we have

$$\begin{aligned} \text{sign}(x^1), \text{sign}(x^2) &\leq \text{sign}(x) \\ \Leftrightarrow \\ \text{sign}\left(\begin{pmatrix} x^1 \\ Ax^1 \end{pmatrix}\right), \text{sign}\left(\begin{pmatrix} x^2 \\ Ax^2 \end{pmatrix}\right) &\leq \text{sign}\left(\begin{pmatrix} x \\ Ax \end{pmatrix}\right). \end{aligned}$$

It remains to show the equivalence of the conclusions in the two definitions. In fact,

$$x^1 = \lambda x^2 \text{ with } \lambda > 0 \Leftrightarrow \begin{pmatrix} x^1 \\ Ax^1 \end{pmatrix} = \lambda \begin{pmatrix} x^2 \\ Ax^2 \end{pmatrix} \text{ with } \lambda > 0. \quad \square$$

Now, we can extend Theorem 3 to general polyhedral cones.

Theorem 8. *Let $C = \{x \mid Ax \geq 0\}$ be a polyhedral cone. Every nonzero vector $x \in C$ is a conformal sum of EVs. That is, there exists a finite set $E \subseteq C$ of EVs such that*

$$x = \sum_{e \in E} e \quad \text{with } \text{sign}(e) \leq \text{sign}(x).$$

The set E can be chosen such that $|E| \leq \dim(C)$ and $|E| \leq |\text{supp}(x)| + |\text{supp}(Ax)|$.

Proof. Let $A \in \mathbb{R}^{m \times r}$. Define the subspace

$$\tilde{S} = \left\{ \begin{pmatrix} x \\ Ax \end{pmatrix} \in \mathbb{R}^{r+m} \mid x \in \text{span}(C) \right\}$$

and the s-cone

$$\tilde{C} = \left\{ \begin{pmatrix} x \\ Ax \end{pmatrix} \in \mathbb{R}^{r+m} \mid x \in C \right\}.$$

Let $x \in C$ be nonzero. By Theorem 3, $\begin{pmatrix} x \\ Ax \end{pmatrix} \in \tilde{C}$ is a conformal sum of EVs. That is, there exists a finite set $\tilde{E} \subseteq \tilde{C}$ of EVs such that

$$\begin{pmatrix} x \\ Ax \end{pmatrix} = \sum_{\begin{pmatrix} e \\ Ae \end{pmatrix} \in \tilde{E}} \begin{pmatrix} e \\ Ae \end{pmatrix} \quad \text{with } \text{sign}\left(\begin{pmatrix} e \\ Ae \end{pmatrix}\right) \leq \text{sign}\left(\begin{pmatrix} x \\ Ax \end{pmatrix}\right).$$

By Lemma 7, the EVs of C and \tilde{C} are in one-to-one correspondence. Hence, there exists a finite set $E = \{e \mid \begin{pmatrix} e \\ Ae \end{pmatrix} \in \tilde{E}\} \subseteq C$ of EVs such that

$$x = \sum_{e \in E} e \quad \text{with } \text{sign}(e) \leq \text{sign}(x).$$

The set \tilde{E} (and hence E) can be chosen such that $|E| = |\tilde{E}| \leq \dim(\tilde{S}) = \dim(C)$ and $|E| = |\tilde{E}| \leq |\text{supp}(\frac{x}{Ax})| = |\text{supp}(x)| + |\text{supp}(Ax)|$. \square

Theorem 8 is a conformal refinement of Minkowski's and Carathéodory's theorems for polyhedral cones. In fact, it remains to show that there are finitely many EVs.

Proposition 9. *For a polyhedral cone, there are finitely many cND vectors up to positive scalar multiples.*

Proof. Let C be a polyhedral cone and \tilde{C} the related s-cone. By Lemma 7, the cND vectors of C and \tilde{C} are in one-to-one correspondence. By Proposition 5, the cND and SM vectors of \tilde{C} coincide, and by Proposition 4, there are finitely many SM vectors. \square

In Urbanczik and Wagner (2005), EVs of a polyhedral cone C were equivalently defined as extreme vectors of intersections of C with closed orthants of maximal dimension. Indeed, the following equivalence holds for closed orthants, not necessarily of maximal dimension.

Proposition 10. *Let $C \subseteq \mathbb{R}^r$ be a polyhedral cone, $x \in C$, and $O \subset \mathbb{R}^r$ a closed orthant with $x \in O$. Then,*

$$x \in C \text{ is cND} \Leftrightarrow x \in C \cap O \text{ is EX.}$$

Proof. We show the equivalence of the premises in the definitions of conformal non-decomposability for C and extremity for $C \cap O$. (The conclusions are identical.) Indeed, assuming $x = x^1 + x^2$, we have

$$\begin{aligned} x^1, x^2 \in C \text{ with } \text{sign}(x^1), \text{sign}(x^2) \leq \text{sign}(x) \\ \Leftrightarrow \\ x^1, x^2 \in C \cap O. \end{aligned}$$

\square Hence,

3.3. Polyhedra

Let P be a polyhedron, that is,

$$P = \{x \in \mathbb{R}^r \mid Ax \geq b\} \quad \text{for some } A \in \mathbb{R}^{m \times r} \text{ and } b \in \mathbb{R}^m.$$

In order to extend Theorem 3 to polyhedra, we introduce corresponding special vectors.

3.3.1. Special Vectors

Let P be a polyhedron. A vector $x \in P$ is called

- a *vertex*, if

$$\begin{aligned} \text{for all } x^1, x^2 \in P \text{ and } 0 < \lambda < 1, \\ x = \lambda x^1 + (1 - \lambda)x^2 \text{ implies } x^1 = x^2, \end{aligned} \quad (\text{VE})$$

- and *convex-conformally non-decomposable*, if

$$\begin{aligned} \text{for all } x^1, x^2 \in P \text{ with } \text{sign}(x^1), \text{sign}(x^2) \leq \text{sign}(x) \text{ and} \\ 0 < \lambda < 1, x = \lambda x^1 + (1 - \lambda)x^2 \text{ implies } x^1 = x^2. \end{aligned} \quad (\text{ccND})$$

From the definitions, we have

$$\text{VE} \Rightarrow \text{ccND}.$$

For a polyhedral cone, we defined elementary vectors (EVs) via conformal non-decomposability. For a polyhedron, we require two sorts of EVs: convex-conformally non-decomposable vectors of the polyhedron and conformally non-decomposable vectors of its recession cone.

Definition 11. *Let $P = \{x \in \mathbb{R}^r \mid Ax \geq b\}$ be a polyhedron and $C^r = \{x \in \mathbb{R}^r \mid Ax \geq 0\}$ its recession cone. A vector $e \in C^r \cup P$ is called an elementary vector of P if either $e \in C^r$ is conformally non-decomposable or $e \in P$ is convex-conformally non-decomposable.*

In order to apply Theorem 3, we define an s-cone related to a polyhedron $P = \{x \in \mathbb{R}^r \mid Ax \geq b\}$. We introduce the *homogenization*

$$C^h = \left\{ \begin{pmatrix} x \\ \xi \end{pmatrix} \in \mathbb{R}^{r+1} \mid \xi \geq 0 \text{ and } Ax - \xi b \geq 0 \right\}$$

of the polyhedron, the subspace

$$\tilde{S} = \left\{ \begin{pmatrix} x \\ \xi \\ Ax - \xi b \end{pmatrix} \in \mathbb{R}^{r+1+m} \mid \begin{pmatrix} x \\ \xi \end{pmatrix} \in \text{span}(C^h) \right\}$$

with $\dim(\tilde{S}) = \dim(C^h) = \dim(P) + 1$, and the s-cone

$$\begin{aligned} \tilde{C} &= C(\tilde{S}, 1 + m) \\ &= \left\{ \begin{pmatrix} x \\ \xi \\ Ax - \xi b \end{pmatrix} \in \mathbb{R}^{r+1+m} \mid \begin{pmatrix} x \\ \xi \end{pmatrix} \in \text{span}(C^h), \xi \geq 0, \text{ and} \right. \\ &\quad \left. Ax - \xi b \geq 0 \right\} \\ &= \left\{ \begin{pmatrix} x \\ \xi \\ Ax - \xi b \end{pmatrix} \in \mathbb{R}^{r+1+m} \mid \begin{pmatrix} x \\ \xi \end{pmatrix} \in C^h \right\}. \end{aligned}$$

Hence,

$$\begin{pmatrix} x \\ \xi \end{pmatrix} \in C^h \Leftrightarrow \begin{pmatrix} x \\ \xi \\ Ax - \xi b \end{pmatrix} \in \tilde{C}.$$

Moreover, the cND vectors of C^r and the ccND vectors of P (as the cND vectors of C^h) are in one-to-one correspondence with the cND vectors of \tilde{C} .

Lemma 12. *Let $P = \{x \mid Ax \geq b\}$ be a polyhedron, $C^r = \{x \mid Ax \geq 0\}$ its recession cone, and*

$$\tilde{C} = \left\{ \begin{pmatrix} x \\ \xi \\ Ax - \xi b \end{pmatrix} \in \mathbb{R}^{r+1+m} \mid \xi \geq 0 \text{ and } Ax - \xi b \geq 0 \right\}$$

the related s-cone. Then,

$$x \in C^r \text{ is cND} \Leftrightarrow \begin{pmatrix} x \\ 0 \\ Ax \end{pmatrix} \in \tilde{C} \text{ is cND}$$

and

$$x \in P \text{ is ccND} \Leftrightarrow \begin{pmatrix} x \\ 1 \\ Ax - b \end{pmatrix} \in \tilde{C} \text{ is cND}.$$

Proof. See Appendix. \square

Now, we can extend Theorem 3 to polyhedra.

Theorem 13. Let $P = \{x \mid Ax \geq b\}$ be a polyhedron and $C^r = \{x \mid Ax \geq 0\}$ its recession cone. Every vector $x \in P$ is a conformal sum of EVs. That is, there exist finite sets $E_0 \subseteq C^r$ and $E_1 \subseteq P$ of EVs such that

$$x = \sum_{e \in E_0} e + \sum_{e \in E_1} \lambda_e e \quad \text{with } \text{sign}(e) \leq \text{sign}(x),$$

$\lambda_e \geq 0$, and $\sum_{e \in E_1} \lambda_e = 1$. (Hence, $|E_1| \geq 1$.)

The set $E = E_0 \cup E_1$ can be chosen such that $|E| \leq \dim(P) + 1$ and $|E| \leq |\text{supp}(x)| + |\text{supp}(Ax)| + 1$.

Proof. By defining an s-cone related to P , applying Theorem 3, and using Lemma 12. See Appendix. \square

Theorem 13 is a conformal refinement of Minkowski's and Carathéodory's theorems for polyhedra. In fact, it remains to show that there are finitely many EVs.

Proposition 14. For a polyhedron, there are finitely many ccND vectors.

Proof. Let P be a polyhedron and \tilde{C} the related s-cone. By Lemma 12, the ccND vectors of P are in one-to-one correspondence with a subset of cND vectors of \tilde{C} . By Proposition 5, the cND and SM vectors of \tilde{C} coincide, and by Proposition 4, there are finitely many SM vectors. \square

EVs of a polyhedron P can be equivalently defined as vertices of intersections of P with closed orthants.

Proposition 15. Let $P \subseteq \mathbb{R}^r$ be a polyhedron, $x \in P$, and $O \subset \mathbb{R}^r$ a closed orthant with $x \in O$. Then,

$$x \in P \text{ is ccND} \iff x \in P \cap O \text{ is VE.}$$

Proof. We show the equivalence of the premises in the definitions of convex-conformal non-decomposability for P and of a vertex for $P \cap O$. (The conclusions are identical.) Indeed, assuming $x = \lambda x^1 + (1 - \lambda)x^2$ with $0 < \lambda < 1$, we have

$$x^1, x^2 \in P \text{ with } \text{sign}(x^1), \text{sign}(x^2) \leq \text{sign}(x)$$

$$\iff$$

$$x^1, x^2 \in P \cap O.$$

We conclude by noting that Theorem 8 is a special case of Theorem 13. If a polyhedron is also a cone, then $P = C^r$, $E_1 = \{0\}$, and $\sum_{e \in E_1} \lambda_e e = 0$. However, we do not use Theorem 8 to prove Theorem 13. In classical proofs of Minkowski's and Carathéodory's theorems, one first studies polyhedral cones and then extends the results to polyhedra by a method called homogenization/dehomogenization; (see e.g., Ziegler, 1995).

3.4. Minimal Generating Sets

For a pointed polyhedral cone, the extreme rays form a minimal set of generators with respect to addition. The set is minimal in the sense that no proper subset forms a generating set and minimal in the even stronger sense that it is contained in every other generating set. Hence, the extreme rays form a *unique* minimal set of generators.

For a general polyhedral cone, there are minimal sets of generators (minimal in the sense that no proper subset forms a generating set), but there is no unique minimal generating set. However, there is a unique minimal set of *conformal* generators, namely the set of elementary vectors.

Recall that elementary vectors of a polyhedral cone are defined as conformally non-decomposable vectors. Indeed, every nonzero element of a polyhedral cone is a conformal sum of elementary vectors (Theorem 8), and every elementary vector is contained in a set of conformal generators.

We make the above argument more formal.

Definition 16. Let C be a polyhedral cone. A subset $G \subseteq C$ is called a conformal generating set if (i) every nonzero vector $x \in C$ is a conformal sum of vectors in G , that is, if there exists a finite set $G_x \subset G$ such that

$$x = \sum_{g \in G_x} g \quad \text{with } \text{sign}(g) \leq \text{sign}(x),$$

and (ii) if $\lambda G = G$ for all $\lambda > 0$.

Proposition 17. Let C be a polyhedral cone, $E \subseteq C$ the set of elementary vectors, and $G \subseteq C$ a conformal generating set. Then, $E \subseteq G$.

Proof. Let $e \in C$ be an elementary vector. Since G is a conformal generating set, we have

$$e = g^* + h \quad \text{with } \text{sign}(g^*), \text{sign}(h) \leq \text{sign}(x),$$

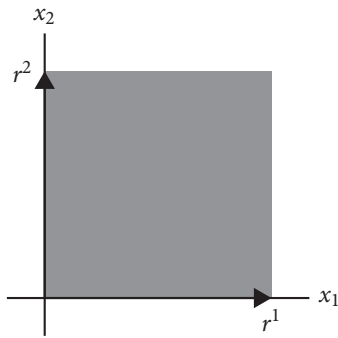
where we choose a nonzero $g^* \in G_e \subset G$ and write $h = \sum_{g \in G_e \setminus \{g^*\}} g \in C$. If $|G_e| = 1$, then $h = 0$ and $e = g^* \in G$. Otherwise, since e is an elementary vector (a cND vector), we have $h = \lambda g^*$ with $\lambda > 0$ and hence $e = (1 + \lambda)g^* \in G$. \square

Analogously, for a polyhedron, there is a unique minimal set of conformal generators, namely the set of elementary vectors.

3.5. Examples

We illustrate our results by examples of polyhedral cones and polyhedra in two dimensions, and we return to the running example from the introduction.

Example 1. The s-cone $C = \{x \mid x_1 \geq 0, x_2 \geq 0\}$.



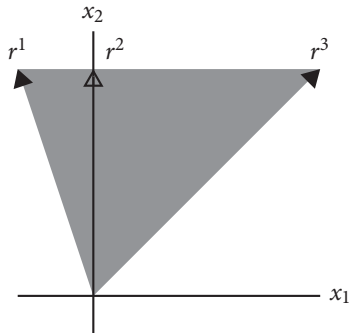
Its EVs (SM vectors) are elements of the rays $r^1 = \{x \mid x_1 > 0, x_2 = 0\}$ and $r^2 = \{x \mid x_1 = 0, x_2 > 0\}$ (indicated by arrows). Every nonzero vector $x \in C$ is a conformal sum of EVs. That is,

$$x = e^1 + e^2,$$

where $e^1 \in r^1$ and $e^2 \in r^2$.

Example 2. The general polyhedral cone

$$C = \{x \mid \begin{pmatrix} 3 & 1 \\ -1 & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \geq 0\}.$$



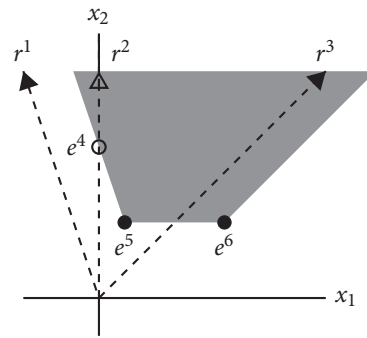
Its EVs (cND vectors) are elements of the rays r^1 , r^2 , and r^3 . Note that r^2 is not an extreme ray of C , but an extreme ray of $C \cap \mathbb{R}_{\geq}^2$, the intersection of the cone with the nonnegative orthant. Every nonzero vector $x \in C$ is a conformal sum of EVs. In particular, if $x \in C \cap \mathbb{R}_{\geq}^2$, then

$$x = e^2 + e^3,$$

where $e^2 \in r^2$ and $e^3 \in r^3$.

Example 3. The polyhedron

$$P = \{x \mid \begin{pmatrix} 3 & 1 \\ -3 & 3 \\ 0 & 2 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \geq \begin{pmatrix} 1 \\ -1 \\ 1 \end{pmatrix}\}.$$

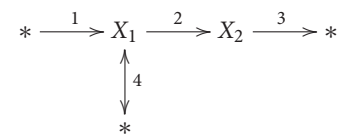


Its EVs are elements of the rays r^1 , r^2 , and r^3 (cND vectors of the recession cone) and the vectors e^4 , e^5 , and e^6 (ccND vectors of the polyhedron). Note that e^4 is not a vertex of P , but a vertex of $P \cap \mathbb{R}_{\geq}^2$, the intersection of the polyhedron with the nonnegative orthant. Every vector $x \in P$ is a conformal sum of EVs. In particular, if $x \in P \cap \mathbb{R}_{\geq}^2$, then

$$x = (e^2 + e^3) + (\lambda_4 e^4 + \lambda_5 e^5 + \lambda_6 e^6),$$

where $e^2 \in r^2$, $e^3 \in r^3$ and $\lambda_4, \lambda_5, \lambda_6 \geq 0$ with $\lambda_4 + \lambda_5 + \lambda_6 = 1$.

Finally, we return to the running example from the introduction. We restate the underlying network, the corresponding stoichiometric matrix and the resulting flux cone:



$$N = \begin{pmatrix} 1 & -1 & 0 & -1 \\ 0 & 1 & -1 & 0 \end{pmatrix},$$

$$C = \{f \in \mathbb{R}^4 \mid Nf = 0 \text{ and } f_1, f_2, f_3 \geq 0\}.$$

Its EVs (SM vectors) are

$$e^1 = \begin{pmatrix} 1 \\ 0 \\ 0 \\ 1 \end{pmatrix}, e^2 = \begin{pmatrix} 0 \\ 1 \\ 1 \\ -1 \end{pmatrix}, e^3 = \begin{pmatrix} 1 \\ 1 \\ 1 \\ 0 \end{pmatrix},$$

and their positive multiples. In other words, the EVs are elements of the rays $r^1 = \{\lambda e^1 \mid \lambda > 0\}$, $r^2 = \{\lambda e^2 \mid \lambda > 0\}$, and $r^3 = \{\lambda e^3 \mid \lambda > 0\}$.

The flux cone is defined by the stoichiometric matrix and the set of irreversible reactions. If additionally lower/upper bounds for the fluxes through certain reactions are known, then one is interested in the resulting flux polyhedron. In the example, we add an upper bound for the flux through reaction 1, in particular, we require $f_1 \leq 2$ and obtain the flux polyhedron

$$P = \{f \in \mathbb{R}^4 \mid Nf = 0, f_1, f_2, f_3 \geq 0, \text{ and } f_1 \leq 2\}.$$

Its EVs are elements of the ray $r^2 = \{\lambda e^2 \mid \lambda > 0\}$ (cND vectors of the recession cone) and the vectors e^1, e^3, e^4 (ccND vectors of the polyhedron), where

$$e^1 = \begin{pmatrix} 2 \\ 0 \\ 0 \\ 2 \end{pmatrix}, e^2 = \begin{pmatrix} 0 \\ 1 \\ 1 \\ -1 \end{pmatrix}, e^3 = \begin{pmatrix} 2 \\ 2 \\ 2 \\ 0 \end{pmatrix}, e^4 = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}.$$

Note that e^3 is not a vertex of P , but a vertex of $P \cap \mathbb{R}_{\geq}^4$, the intersection of the polyhedron with the nonnegative orthant. Every vector $x \in P$ is a conformal sum of EVs. In particular, if $x \in P \cap \mathbb{R}_{\geq}^4$, then

$$x = \lambda_1 e^1 + \lambda_3 e^3 + \lambda_4 e^4,$$

where $\lambda_1, \lambda_3, \lambda_4 \geq 0$ with $\lambda_1 + \lambda_3 + \lambda_4 = 1$. In other words, the polyhedron $P \cap \mathbb{R}_{\geq}^4$ is a polytope.

In applications such as computational strain design, the set of EVs (the unique minimal set of *conformal* generators) is often more useful than a minimal set of generators. In the example, the set of EVs includes e^3 which is a ccND vector, but not a vertex of P . If we delete reaction 4 by gene knockout, the new set of EVs consists of e^3 and e^4 (having zero flux through reaction 4), and the resulting flux polyhedron is the polytope generated by e^3 and e^4 . Most importantly, we obtain the result without recalculating the set of generators (after deleting reaction 4).

4. DISCUSSION

Metabolic pathway analysis aims to identify meaningful routes in a network, in particular, to decompose fluxes into *minimal* metabolic pathways. However, only a decomposition *without cancelations* is biochemically meaningful, since a reversible reaction cannot have a flux in different directions at the same time.

In mathematical terms, one is interested in a *conformal* decomposition of the flux cone and of general polyhedral cones and polyhedra. In this work, we first study s-cones (like the flux cone) arising from a linear subspace and nonnegativity conditions. Then, we analyze general polyhedral cones and polyhedra via corresponding higher-dimensional s-cones. Without assuming previous knowledge of polyhedral geometry, we provide an elementary proof of a conformal refinement of Minkowski's and Carathéodory's theorems (Theorems 3, 8, and 13): Every vector (of an s-cone, a general polyhedral cone, or a polyhedron) is a conformal sum of *elementary vectors* (conformally non-decomposable vectors), and there is an upper bound on the number of elementary vectors needed in a conformal decomposition (in terms of the dimension of the cone or polyhedron).

As a natural next question, one may ask: what is a *minimal generating set* of a polyhedral cone that allows a conformal

decomposition of every vector? Clearly, such a set must contain all conformally non-decomposable vectors. Indeed, we show that the elementary vectors form a *unique* minimal set of conformal generators (Proposition 17). In metabolic pathway analysis, the question is: what is a minimal generating set of the flux cone that allows a biochemically meaningful decomposition of every flux mode? In this case, the *elementary modes* form a unique minimal set of generators without cancelations. This property distinguishes elementary modes as a fundamental concept in metabolic pathway analysis and may serve as a definition.

The correspondence of general polyhedral cones and polyhedra to higher-dimensional s-cones has also important consequences for the *computation* of elementary vectors. In particular, it allows to use efficient algorithms and software developed for elementary modes (see e.g., Zanghellini et al., 2013 and the references therein) for computing elementary vectors of general polyhedral cones and polyhedra.

In *applications*, decompositions without cancelations were first used in the study of the conversion cone Urbanczik and Wagner (2005), a general polyhedral cone obtained by flux cone projection Marashi et al. (2012). The approach was extended to polyhedra arising from the flux cone and inhomogeneous constraints, in particular, to describe the solution set of linear optimization problems encountered in flux balance analysis Urbanczik (2007). In analogy to s-cones, these sets could be called s-polyhedra. Recently, elementary vectors have been used to describe such polyhedra in the study of growth-coupled product synthesis Klamt and Mahadevan (2015). Interestingly, conformal decompositions of the flux cone itself appeared rather late. In fact, they have been used to characterize optimal solutions of enzyme allocation problems in *kinetic* metabolic networks Müller et al. (2014).

Minkowski's and Carathéodory's theorems (and their conformal refinements) are fundamental results in polyhedral geometry with important applications in metabolic pathway analysis. In subsequent work, we plan to revisit other results from polyhedral geometry and oriented matroids (like Farkas' lemma) and investigate their consequences for metabolic pathway analysis.

AUTHOR CONTRIBUTIONS

The authors contributed equally to the work and approved it for publication.

ACKNOWLEDGMENTS

The authors acknowledge helpful comments from three reviewers. SM was supported by the Austrian Science Fund (FWF), project P28406. GR was supported by the FWF, project P27229.

REFERENCES

- Bachem, A., and Kern, W. (1992). *Linear Programming Duality. An Introduction to Oriented Matroids*. Berlin: Springer-Verlag.
- Carathéodory, C. (1911). Über den Variabilitätsbereich der Fourierschen Konstanten von positiven harmonischen Funktionen. *Rend. Circ. Mat. Palermo*, 32, 193–217.
- Klamt, S., and Mahadevan, R. (2015). On the feasibility of growth-coupled product synthesis in microbial strains. *Metab. Eng.* 30, 166–178. doi: 10.1016/j.ymben.2015.05.006
- Klamt, S., and Stelling, J. (2003). Two approaches for metabolic pathway analysis? *Trends Biotechnol.* 21, 64–69. doi: 10.1016/S0167-7799(02)00034-3
- Llaneras, F., and Picó, J. (2010). Which metabolic pathways generate and characterize the flux space? A comparison among elementary modes, extreme pathways and minimal generators. *J. Biomed. Biotechnol.* 2010:753904. doi: 10.1155/2010/753904
- Marashi, S.-A., David, L., and Bockmayr, A. (2012). Analysis of metabolic subnetworks by flux cone projection. *Algorithms Mol. Biol.* 7:17. doi: 10.1186/1748-7188-7-17
- Minkowski, H. (1896). *Geometrie der Zahlen (Erste Lieferung)*. Teubner, Leipzig.
- Müller, S., Regensburger, G., and Steuer, R. (2014). Enzyme allocation problems in kinetic metabolic networks: optimal solutions are elementary flux modes. *J. Theor. Biol.* 347, 182–190. doi: 10.1016/j.jtbi.2013.11.015
- Rockafellar, R. T. (1969). “The elementary vectors of a subspace of \mathbb{R}^N ,” in *Combinatorial Mathematics and its Applications (Proc. Conf., Univ. North Carolina, Chapel Hill, N.C., 1967)* (Chapel Hill, NC: Univ. North Carolina Press), 104–127.
- Schuster, S. (2015). Personal communication at MPA 2015, Braga, Portugal.
- Schuster, S., and Hilgetag, C. (1994). On elementary flux modes in biochemical reaction systems at steady state. *J. Biol. Syst.* 2, 165–182.
- Schuster, S., Hilgetag, C., Woods, J. H., and Fell, D. A. (2002). Reaction routes in biochemical reaction systems: algebraic properties, validated calculation procedure and example from nucleotide metabolism. *J. Math. Biol.* 45, 153–181. doi: 10.1007/s002850200143
- Urbanczik, R. (2007). Enumerating constrained elementary flux vectors of metabolic networks. *IET Syst. Biol.* 1, 274–279. doi: 10.1049/iet-syb:20060073
- Urbanczik, R., and Wagner, C. (2005). Functional stoichiometric analysis of metabolic networks. *Bioinformatics* 21, 4176–4180. doi: 10.1093/bioinformatics/bti674
- Zanghellini, J., Ruckerbauer, D. E., Hanscho, M., and Jungreuthmayer, C. (2013). Elementary flux modes in a nutshell: properties, calculation and applications. *Biotechnol. J.* 8, 1009–1016. doi: 10.1002/biot.201200269
- Ziegler, G. M. (1995). *Lectures on Polytopes*. New York, NY: Springer-Verlag.

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2016 Müller and Regensburger. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

APPENDIX

We prove the main results for polyhedra, Lemma 12 and Theorem 13.

Proof of Lemma 12. To prove the first equivalence, we note that $\begin{pmatrix} x \\ 0 \\ Ax \end{pmatrix} \in \tilde{C}$ is cND if and only if $\begin{pmatrix} x \\ Ax \end{pmatrix} \in C'$ is cND, where $C' = \{\begin{pmatrix} x \\ Ax \end{pmatrix} \in \mathbb{R}^{r+m} \mid Ax \geq 0\}$, and apply Lemma 7.

To prove the second equivalence, we show the two implications separately:

(\Rightarrow) We assume that $x \in P$ is ccND and first consider a conformal sum of the form

$$\begin{pmatrix} x \\ 1 \\ Ax-b \end{pmatrix} = \begin{pmatrix} x^1 \\ 1 \\ Ax^1-b \end{pmatrix} + \begin{pmatrix} x^2 \\ 0 \\ Ax^2 \end{pmatrix}$$

with $x^1 \in P$, nonzero $x^2 \in C^r$, and $\text{sign}(x^1), \text{sign}(x^2) \leq \text{sign}(x)$. Indeed, we also have $x = \frac{1}{2}x^1 + \frac{1}{2}(x^1 + 2x^2)$ with $x^1, x^1 + 2x^2 \in P$ and $\text{sign}(x^1), \text{sign}(x^1 + 2x^2) \leq \text{sign}(x)$. By the assumption, $x^1 = x^1 + 2x^2$, that is, $x^2 = 0$, and it remains to consider a conformal sum of the form

$$\begin{pmatrix} x \\ 1 \\ Ax-b \end{pmatrix} = \lambda \begin{pmatrix} x^1 \\ 1 \\ Ax^1-b \end{pmatrix} + (1-\lambda) \begin{pmatrix} x^2 \\ 1 \\ Ax^2-b \end{pmatrix} \quad (+)$$

with $x^1, x^2 \in P$, $\text{sign}(x^1), \text{sign}(x^2) \leq \text{sign}(x)$, and $0 < \lambda < 1$. By the assumption, $x^1 = x^2$, and the first vector in the sum is a positive multiple of the second. That is,

$$\lambda \begin{pmatrix} x^1 \\ 1 \\ Ax^1-b \end{pmatrix} = \mu (1-\lambda) \begin{pmatrix} x^2 \\ 1 \\ Ax^2-b \end{pmatrix} \quad (*)$$

with $\mu > 0$. Hence, $\begin{pmatrix} x \\ 1 \\ Ax-b \end{pmatrix} \in \tilde{C}$ is cND.

(\Leftarrow) We assume that $\begin{pmatrix} x \\ 1 \\ Ax-b \end{pmatrix} \in \tilde{C}$ is cND and consider the convex-conformal sum

$$x = \lambda x^1 + (1-\lambda)x^2$$

with $x^1, x^2 \in P$, $\text{sign}(x^1), \text{sign}(x^2) \leq \text{sign}(x)$, and $0 < \lambda < 1$. Hence, we also have the conformal sum (+). By the assumption,

we have equation (*) which implies $x^1 = x^2$. Hence, $x \in P$ is ccND. \square

Proof of Theorem 13. Let $A \in \mathbb{R}^{m \times r}$ and $b \in \mathbb{R}^m$. Define the homogenization

$$C^h = \{\begin{pmatrix} x \\ \xi \end{pmatrix} \in \mathbb{R}^{r+1} \mid \xi \geq 0 \text{ and } Ax - \xi b \geq 0\},$$

the subspace

$$\tilde{S} = \{\begin{pmatrix} x \\ \xi \\ Ax-\xi b \end{pmatrix} \in \mathbb{R}^{r+1+m} \mid \begin{pmatrix} x \\ \xi \end{pmatrix} \in \text{span}(C^h)\}$$

and the s-cone

$$\tilde{C} = \{\begin{pmatrix} x \\ \xi \\ Ax-\xi b \end{pmatrix} \in \mathbb{R}^{r+1+m} \mid \begin{pmatrix} x \\ \xi \end{pmatrix} \in C^h\}.$$

Let $x \in P$. By Theorem 3, $\begin{pmatrix} x \\ 1 \\ Ax-b \end{pmatrix} \in \tilde{C}$ is a conformal sum of EVs. That is, there exist finite sets $\tilde{E}_0, \tilde{E}_1 \subseteq \tilde{C}$ of (normalized) EVs such that

$$\begin{pmatrix} x \\ 1 \\ Ax-b \end{pmatrix} = \sum_{\begin{pmatrix} e \\ 0 \\ Ae \end{pmatrix} \in \tilde{E}_0} \begin{pmatrix} e \\ 0 \\ Ae \end{pmatrix} + \sum_{\begin{pmatrix} e \\ 1 \\ Ae-b \end{pmatrix} \in \tilde{E}_1} \lambda_e \begin{pmatrix} e \\ 1 \\ Ae-b \end{pmatrix}$$

with

$$\text{sign}\begin{pmatrix} e \\ 0 \\ Ae \end{pmatrix}, \text{sign}\begin{pmatrix} e \\ 1 \\ Ae-b \end{pmatrix} \leq \text{sign}\begin{pmatrix} x \\ 1 \\ Ax-b \end{pmatrix},$$

$\lambda_e \geq 0$, and $\sum_{e \in E_1} \lambda_e = 1$. By Lemma 12, the EVs of P are in one-to-one correspondence with the EVs of \tilde{C} . Hence, there exist finite sets $E_0 = \{e \mid \begin{pmatrix} e \\ 0 \\ Ae \end{pmatrix} \in \tilde{E}_0\} \subseteq C^r$ and $E_1 = \{e \mid \begin{pmatrix} e \\ 1 \\ Ae-b \end{pmatrix} \in \tilde{E}_1\} \subseteq P$ of EVs such that

$$x = \sum_{e \in E_0} e + \sum_{e \in E_1} \lambda_e e \quad \text{with } \text{sign}(e) \leq \text{sign}(x).$$

The set $\tilde{E} = \tilde{E}_0 \cup \tilde{E}_1$ (and hence $E = E_0 \cup E_1$) can be chosen such that $|\tilde{E}| = |\tilde{E}_1| \leq \dim(\tilde{S}) = \dim(P) + 1$ and $|E| = |\tilde{E}| \leq |\text{supp}\begin{pmatrix} x \\ 1 \\ Ax-b \end{pmatrix}| = |\text{supp}(x)| + 1 + |\text{supp}(Ax - b)|$. \square



Identification of Metabolic Pathway Systems

Sepideh Dolatshahi[†] and Eberhard O. Voit^{*}

Department of Biomedical Engineering, Georgia Institute of Technology, Atlanta, GA, USA

OPEN ACCESS

Edited by:

Rui Alves,
Universitat de Lleida, Spain

Reviewed by:

Osbaldo Resendis-Antonio,
Instituto Nacional De Medicina
Genómica, Mexico
Satyaprakash Nayak,
Pfizer Inc., USA
Nestor V. Torres,
Universidad de La Laguna, Spain
Armando José Salvador,
Center for Neuroscience and Cell
Biology, Portugal

*Correspondence:

Eberhard O. Voit
eberhard.voit@bme.gatech.edu

[†]Present Address:

Sepideh Dolatshahi,
Penn Institute for Immunology,
University of Pennsylvania,
Philadelphia, PA, USA

Specialty section:

This article was submitted to
Systems Biology,
a section of the journal
Frontiers in Genetics

Received: 09 October 2015

Accepted: 18 January 2016

Published: 10 February 2016

Citation:

Dolatshahi S and Voit EO (2016)
Identification of Metabolic Pathway
Systems. *Front. Genet.* 7:6.
doi: 10.3389/fgene.2016.00006

The estimation of parameters in even moderately large biological systems is a significant challenge. This challenge is greatly exacerbated if the mathematical formats of appropriate process descriptions are unknown. To address this challenge, the method of dynamic flux estimation (DFE) was proposed for the analysis of metabolic time series data. Under ideal conditions, the first phase of DFE yields numerical representations of all fluxes within a metabolic pathway system, either as values at each time point or as plots against their substrates and modulators. However, this numerical result does not reveal the mathematical format of each flux. Thus, the second phase of DFE selects functional formats that are consistent with the numerical trends obtained from the first phase. While greatly facilitating metabolic data analysis, DFE is only directly applicable if the pathway system contains as many dependent variables as fluxes. Because most actual systems contain more fluxes than metabolite pools, this requirement is seldom satisfied. Auxiliary methods have been proposed to alleviate this issue, but they are not general. Here we propose strategies that extend DFE toward general, slightly underdetermined pathway systems.

Keywords: dynamic flux estimation (DFE), identifiability, metabolic pathway analysis, parameter estimation, pathway structure, underdetermined system of fluxes

INTRODUCTION AND BACKGROUND

A Google Scholar search for the keyword “parameter estimation” yields over 3 million hits, which renders it abundantly evident that the topic is everything but trivial, especially for applications in biology. The challenges of finding optimal parameter values for biological systems are multifold and include mathematical, statistical, computational, and even biological aspects. Mathematical issues include dependencies among parameter values, sloppiness, and different types of exact or approximate compensation between errors among the equations of the system, within equations, and even within terms of the equations. Computational challenges are driven by the sheer size of the often high-dimensional parameter space, the need to solve systems of differential equations thousands of times, and an error structure between model results and biological data that can be incredibly rough and contain uncounted local minima where search algorithms can get trapped. Biological issues include the size and complexity of a system, noisy or missing data, ill-characterized processes, and unrealistic parameter values. All these challenges are tightly interwoven and often create situations where no (good) solutions are obtained, where too many possible solutions can be identified, or where the exclusive criterion of the quality of the fit is misleading.

Partial help for overcoming some of these complications was provided by the insight that systems of ordinary differential equations (ODEs) can be estimated in a much simplified manner, at least to some degree. Namely, if data are available as time series measurements, and if it is possible

to estimate the slopes of these time courses with some reliability, then the derivatives on the left-hand sides of the ODEs can be replaced with estimated slopes at many time points (Varah, 1982; Voit and Savageau, 1982a,b; Voit, 2000; Voit and Almeida, 2004; Chou and Voit, 2009; Jia et al., 2011). Consequently, each ODE, evaluated at a set of time points, is replaced with a purely algebraic system of equations, where the fluxes constitute its unknown variables. Each of these sets can be evaluated independently of all other sets and does no longer require numerical integration, which can account for more than 95% of the computational cost when parameters are directly estimated for ODEs (Voit and Almeida, 2004). The initial estimation of slopes from the time course data can be accomplished with a variety of methods that range from primitive to sophisticated (e.g., see Whittaker, 1923; Voit and Savageau, 1982b; Eilers, 2003; Voit and Almeida, 2004; Vilela, 2007, 2008; Dolatshahi et al., 2014 and discussions therein).

While it certainly simplifies parameter estimation, the slope estimation and decoupling method is not without its own issues. In particular, it may “warp” solutions in the direction of time, so that, for instance, oscillations have a predicted frequency that is too high or too low (see Chapter 5 of Voit, 2012). Nonetheless, the method can serve as an effective first stab at a complicated problem and thereby provide reasonable initial guesses for standard estimation techniques.

A prerequisite for any parameter estimation effort is knowledge of the mathematical formats of all involved processes, or at least a set of reasonable assumptions regarding these formats, because they obviously dictate the role of each parameter. However, guidelines regarding optimal formats for biological process descriptions are not provided by nature. Linear functions have been very successful in engineering, but it has become clear that they are inadequate for representing many biological phenomena. Thus, one needs to resort to non-linear representations, of which, of course, there are infinitely many. One could argue that biological systems must satisfy the laws of physics, but it is usually impossible to deconvolve biological processes neatly into physical components that can be represented based on physical theory (Voit et al., 2010; Voit, 2013a). To circumvent this problem, many biological systems modelers tend to use certain default representations that have a justification in specific, and often simplified instances but do certainly not tell the whole truth about a biological system *in vivo* or are valid in other contexts (Voit et al., 2015). Arguably the best studied example is the Michaelis–Menten rate law, which is approximately true in carefully crafted experiments *in vitro*, but whose prerequisites are most certainly violated in actual biological systems *in situ* (Savageau, 1992, 1995). Similarly, mass action functions in biochemistry, SIR models in epidemiology, and Lotka–Volterra models in ecology may be excellent starting points for the design of models, but it is quite evident that they cannot truly capture the full complexity of living systems in all its details.

One might think that it does not matter too much if the functional form is not perfect, as long as all data of interest are fit with sufficient accuracy. This argument may be true if future predictions and explanations only pertain to the data ranges used

for model parameterization. However, as soon as the model is extrapolated into new ranges of its state variables, extrapolations with the wrong model may lead to grossly unsatisfactory results (Goel et al., 2008). One root cause of such extrapolation problems is a compensation of errors, which may occur within fluxes, among fluxes of the same equation, and among fluxes of different equations. While such compensation can lead to acceptable residual errors in the original data fit, extrapolations to new conditions can become rather unreliable; for specific details see Supplements of Goel et al. (2008).

Faced with this conundrum, the method of dynamic flux estimation (DFE) was suggested for the analysis of metabolic time series data (Goel et al., 2008). In principle, DFE could be applicable to any types of ODE systems, such as gene regulatory networks that offer similar identification challenges (Siegenthaler and Gunawan, 2014; Ud-Dean and Gunawan, 2014), but a very beneficial feature of metabolic systems is the conservation of mass at each metabolite pool, which has as a consequence that many fluxes appear in more than one equation. It will become evident throughout this article that this fact is important for DFE.

DFE consists of two phases, the first of which is model-free and makes very few assumptions (Figure 1). It includes data preprocessing, time course smoothing, the estimation of slopes of the smoothed time courses, and the solution of linear algebraic systems. Generically, each equation of the ODE is written as

$$\frac{dX_i}{dt} = \text{Influx}_1^i + \text{Influx}_2^i + \text{Influx}_3^i + \dots - \text{Efflux}_1^i - \text{Efflux}_2^i - \dots \quad (1)$$

At each time point, the left-hand side is replaced by the appropriate slope, and the equations are simultaneously valid for all time points. The ultimate result of this phase consists of numerical or graphical time series profiles of all fluxes; in other words, the analysis yields plots of the fluxes in the system against time or against metabolites and modulators. Importantly, this phase does not reveal functional formats (Figure 2).

The second phase of DFE is dedicated to the mathematical characterization and parameterization of each flux profile. This phase requires the assumption of functional formats, which are fitted against the numerical flux representations. This step requires parameter estimation, but it is much simpler than the estimation of the original ODE systems, because it now targets explicit functions of one or a few variables in isolation and with correspondingly few parameters. For instance, the graphical result in Figure 2A might suggest a Hill or logistic function as a reasonable format, while appropriate formats for the trends in Figures 2B,C are not clear. It should be noted that this estimation of individual fluxes avoids many of the error compensation issues mentioned before.

Once a mathematical format is chosen for a particular flux, the data are fitted against this alleged format or against a roster of candidate functional forms. No generic strategies exist at this point for selecting candidates or proving their optimality, and it might be useful to scan through a list of candidate functions; for a similar approach in statistics, see Sorribas et al. (2000). Within this list, one may then attempt to identify the best fitting

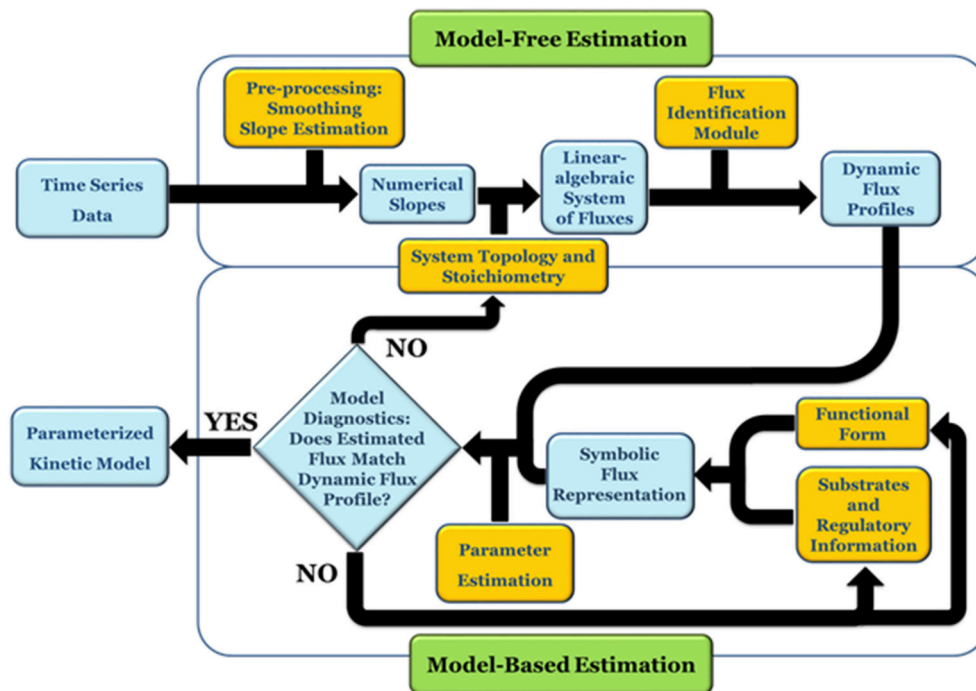


FIGURE 1 | DFE is a model characterization strategy and consists of two phases (adapted from Goel et al., 2008). In the first, model-free estimation phase, it takes time series of concentration data as input and estimates the dynamic flux profiles, which in turn are used as input to phase 2, which consists of a model-based estimation. In this phase, functional forms and regulatory assumptions are incorporated and parameters are estimated for each flux separately.

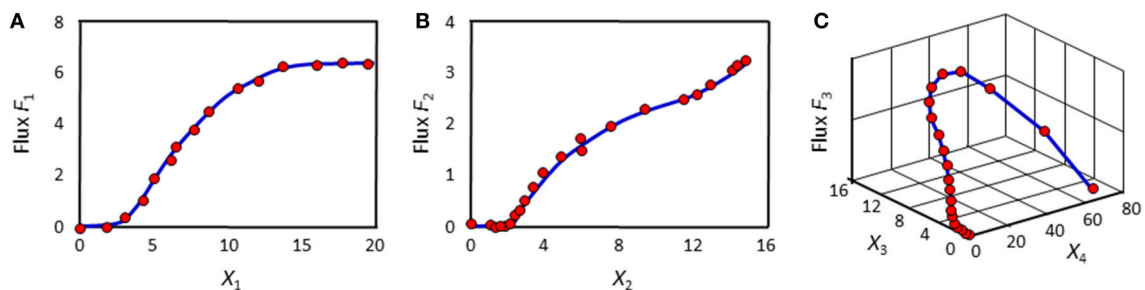


FIGURE 2 | Example results of the first phase of DFE. The flux in panel (A) might be representable with a Hill or logistic function. The fluxes in panels (B) and (C) are adapted from Chou and Voit (2012); their optimal mathematical formats are unknown.

format through regression diagnostics, such as the residual error and a *run test for residuals* (Draper and Smith, 1981). The special case of the power-law format simplifies this step (Savageau and Voit, 1982), as a logarithmic transformation yields linearity and thus permits testing of the appropriateness of a functional form with diagnostic methods of multiple linear regression, even though one has to consider the distortion of the error structure due to the transformation. It is possible that several candidate functions are equally plausible and lead to similar fits. For instance, a Hill function and a logistic function can have essentially indistinguishable graphs. It is also possible that no functional form may be capable of yielding a reasonable fit, which may suggest the existence of missing features in the models,

such as regulatory signals that had not been taken into account in the assumed pathway structure. Such suggestions correspond to novel hypotheses that are testable with further experiments and may lead to biological discoveries, as was demonstrated in Dolatshahi et al. (2016a).

The first phase of DFE mandates that an algebraic system of fluxes be solved at each time point (see Equation 1). This process is straightforward if the number of independent fluxes equals the number of dependent variables for which data exist. However, if the stoichiometric matrix of the system is not full-rank, which actually is the most common case, a direct inversion is not possible, and one needs to resort to auxiliary methods or mathematical operations that cast the problem in a simpler

form (Jia et al., 2012; Liu and Gunawan, 2014). Unfortunately, such methods often necessitate additional biological information to make the stoichiometric matrix invertible (e.g., Voit, 2009; Chou and Voit, 2012; Iwata et al., 2013). As a consequence, these methods are seldom general and often require specific features of the data.

As an alternative or complementation of these methods, this article describes a generic flux identification procedure for slightly underdetermined systems and characterizes the space of available fluxes. The article furthermore discusses conceptual strategies for dealing with missing data and proposes mixed parameter estimation strategies when DFE is only partially applicable. This section involves the second, model-based phase of DFE.

In reality, biological data are always noisy and often incomplete, which adds uncertainty to any estimation or identification method. Indeed, noise, missing data, and estimation issues lead to a complicated intermixing of errors that are difficult to deconvolve. In order to focus exclusively on issues directly associated with the identification of fluxes, we decided here to use “ideal” data, which we generated with a published model (Curien et al., 2009). Many authors have discussed means of addressing and smoothing noisy data and dealing with less than ideal data (e.g., Vilela, 2007; Voit, 2011; Dolatshahi et al., 2014; and references therein), so that we will not revisit this issue here. However, we note that methods very similar to those presented here were recently applied to an actual, rather complex system (Dolatshahi et al., 2016a,b).

CHARACTERIZATION OF METABOLIC FLUXES FROM TIME SERIES DATA

If a pathway system is underdetermined, DFE cannot directly be applied. The issue in this case is not the absence of a solution; rather, the challenge is the existence of an entire space of feasible solutions and the need to decide which of these solutions are in some sense “better” than others. One could explore whether certain normalization or regularization procedures might help, but it appears that they do not solve the problem here, as we simply do not know what type of flux distribution nature considers optimal. For instance, the use of the Moore-Penrose pseudo-inverse (Penrose, 1955; Albert, 1972) yields a solution, but some fluxes of this solution are typically negative, which is often biologically unrealistic. Characterizability analysis (Voit, 2013b) reveals which fluxes within an underdetermined system can be estimated with DFE without additional information, but does not suggest further steps toward an optimal solution. The strategy of the following sections will be to study the entire set of feasible solutions in a drastically reduced space, whose dimension equals the number of the degrees of freedom within the stoichiometric system.

Along with the exploration of the solution space, useful strategies will be introduced to visualize feasible candidate sets. Initially, no information about the functional forms and the contributing metabolites and modulators of each flux is assumed to be available. Later on, minimal generic features of metabolic

fluxes are suggested as constraints to improve the results. It is noted, though, that, even with these constraints, the solutions are not necessarily unique. Finally, solutions in the form of point-wise numerically defined fluxes will be suggested that are appropriate, if not optimal, according to certain criteria of biological reasonableness.

The source code for the following analyses has been deposited on github (<https://github.com/sepid/Identification-of-Metabolic-Pathway-Systems>) and is also presented in the Supplementary Material.

Mathematical Formulation of the Problem

A metabolic pathway system as formulated in Equation (1) can be written in general matrix and vector notation as

$$\frac{d\mathbf{X}}{dt} = \dot{\mathbf{X}} = \mathbf{A}\mathbf{v}. \quad (2)$$

Here, \mathbf{X} denotes a vector of n metabolite concentrations and \mathbf{v} is a vector of m fluxes, i.e., reaction rates, while \mathbf{A} is the stoichiometric matrix. The vectors, but not the matrix, change with time, and the functional forms governing the fluxes are functions of their substrates and regulators. They are in general unknown or based on assumptions that might or might not hold under the given experimental conditions or *in vivo*. Moreover, in certain cases, regulators and cofactors are yet to be discovered and are therefore falsely omitted. This uncertainty is the reason to attempt minimizing assumptions while executing the task of inferring flux profiles from metabolic time series data. At the same time, DFE provides us in this phase with the option of testing and challenging some of the prior assumptions and possibly discovering missing regulatory effects (cf. Dolatshahi et al., 2016a).

Assuming that data smoothing and slope estimation had been conducted successfully at each time point t_i , we replace the left-hand side of Equation (2) with the vector of slopes at time t_i , which we call $\mathbf{b}(t_i)$. Equation (2) can thus be written as a set of algebraic equations. Specifically, suppose that $\mathbf{b}(t) = [\dot{X}_1(t), \dots, \dot{X}_n(t)]^T$ is the vector of slopes of dependent variables at time t and \mathbf{A} is the $n \times m$ stoichiometric matrix, which is constant throughout the time period of any given experiment. Then we obtain directly the linear algebraic system

$$\mathbf{A}\mathbf{v}(t) = \mathbf{b}(t) \quad (3)$$

At a steady state, or when the numerical values of the derivatives are known, Equation (3) has a solution that can be computed for every time point by matrix inversion, if the system has full rank. However, most metabolic systems are under-determined, so that a unique solution does not exist.

We can thus distinguish three situations. (1) When the system has maximal rank, the solution is obtained with the regular inverse, so that $\mathbf{v}(t_i) = \mathbf{A}^{-1} \mathbf{b}(t_i)$ is the solution of the system of equations. (2) When the system is over-determined and has more equations than unknowns ($m < n$), the Moore-Penrose pseudo-inverse \mathbf{A}^+ of matrix \mathbf{A} minimizes the sum of squared errors, $\arg \min(\|\mathbf{A}\mathbf{v}(t_i) - \mathbf{b}(t_i)\|) = \mathbf{A}^+ \mathbf{b}(t_i)$. This

solution is equivalent to the result of linear regression. Finally, (3), the case of under-determined systems ($m > n$) is the most common situation in metabolic modeling, because most pathway systems contain more reaction steps than metabolites. This common occurrence makes the under-determined case particularly important for the model-free phase of DFE and suggests that we investigate if the pseudo-inverse solution $\mathbf{v}(t_i) = A^+ \mathbf{b}(t_i)$ constitutes a biologically feasible, or even optimal, solution.

Pseudo-inverses have been used to solve under-determined systems for a long time. They are characterized by the minimum L^2 -norm within a one- or higher-dimensional space of admissible solutions, i.e., $\arg \min(\|A\mathbf{v}(t_i) - \mathbf{b}(t_i)\|)$. While the best solution, in terms of the smallest norm, is guaranteed by the pseudo-inverse, the resulting fluxes are not necessarily positive, and there is no guarantee that they are biologically meaningful, let alone optimal. In fact, experience shows that minimum-norm solutions often include negative values, which are not biologically feasible as flux values, unless one permits flux inversion, which is not always realistic. The issue of under-determined systems in DFE has been known since the inception of the method, and characterizability analysis, based on pseudo-inverses, was introduced as an *a priori*, data-independent check for the applicability of DFE given a particular pathway system (Voit, 2013b).

A Compact Representation: Gamma-Space and Gamma-Trajectory

In order to characterize the space of admissible flux sets $\mathbf{v}(t) = [v_1(t), \dots, v_m(t)]^T$ $t \in [0, \infty)$ in an efficient manner, a more compact representation is desirable. For pathways with m fluxes and n dependent variables, where $m > n$, let d be the number of degrees of freedom (DOF): $d \geq m - n$. Without loss of generality, we assume that the rank of the system is n . At each time point t , the space of solutions satisfying Equation (3) can be written as:

$$\mathbf{v}(t) = A^+ \mathbf{b}(t) + (\mathbb{I}_m - A^+ A) \mathbf{w}(t) = A^+ \mathbf{b}(t) + \text{null}(A) \boldsymbol{\gamma}(t) \quad (4)$$

Here, $A^+ = A^T(AA^T)^{-1}$ is the Moore-Penrose pseudo-inverse, $A^+ \mathbf{b}(t)$ is the minimum-norm flux set at time t , and \mathbb{I}_m is the $m \times m$ identity matrix. While $A^+ \mathbf{b}(t)$ is easily computed for practical applications with software like MATLAB, the result often contains one or more negative fluxes for some time points, which is usually undesirable. However, if $\mathbf{w}(t_i)$ is a vector of m arbitrary, real-valued elements, then the complete solution $\mathbf{v}(t_i) = A^+ \mathbf{b}(t_i) + (\mathbb{I}_m - A^+ A) \mathbf{w}(t_i)$ represents all possible solutions and spans the null space of the stoichiometric matrix A . In numerical evaluations, this null space is readily determined with the `null(A)` command in MATLAB.

The columns of $\text{null}(A) = [\text{vec}_1, \text{vec}_2, \dots, \text{vec}_d]$ span the null space of A , and $\boldsymbol{\gamma}(t) = [\gamma_1(t), \gamma_2(t), \dots, \gamma_d(t)]^T$ is the corresponding vector of coefficients at time t . Each feasible solution of Equation (3) at time t can thus be uniquely represented by $\boldsymbol{\gamma}(t)$. This representation allows us to explore the d -dimensional *Gamma-space* instead of the feasible subset of the

m -dimensional space of fluxes, whose visual representation is much more challenging.

The representations for all time points are now collected as follows. For each time point t , a feasible flux set $\mathbf{v}(t)$ can be calculated by finding Gamma coefficients that satisfy $\mathbf{v}_{\text{null}}(t) = \text{null}(A) \boldsymbol{\gamma}(t) = [v_1(t), \dots, v_m(t)]^T - A^+ \mathbf{b}(t)$. This equation can be assessed by projecting $\mathbf{v}_{\text{null}}(t)$ onto the vectors $\text{vec}_1, \text{vec}_2, \dots, \text{vec}_d$, which span the null space of A . The coefficient vector $[\gamma_1(t), \dots, \gamma_d(t)]$ constitutes a point in the d -dimensional Gamma-space, representing time point t , and the collection of these points constitutes a trajectory, which we call the *Gamma-trajectory*. Each Gamma-trajectory uniquely represents a feasible flux set traversing all time points, as long as this trajectory corresponds exclusively to non-negative fluxes.

As an illustration, let us consider a simple network consisting of two dependent variables and four fluxes (Figure 3A). Its stoichiometric representation is

$$\begin{bmatrix} 1 & -1 & 0 & 0 \\ 0 & 1 & -1 & -1 \end{bmatrix} \begin{bmatrix} v_1(t) \\ v_2(t) \\ v_3(t) \\ v_4(t) \end{bmatrix} = \begin{bmatrix} b_1(t) \\ b_2(t) \end{bmatrix} \quad (5)$$

Suppose that metabolite concentrations $X_1(t)$ and $X_2(t)$ have been measured every 30 s between 0 and 15 min. Finding the slopes of the concentration trends directly yields $b_1(t)$ and $b_2(t)$ (Figure 3B). The feasible space of solutions, in terms of fluxes, is a two-dimensional plane within a 4-dimensional space that is difficult to visualize directly. Figure 3C shows some representative flux solutions. Even though these are very different, and several of them in fact have little similarity to the fluxes in the model used to generate the “data” (black curves in Figure 3C), all these fluxes satisfy Equation (5) exactly. The corresponding Gamma-trajectories are depicted in Figure 3D. The fluxes and Gamma-trajectory with which the concentration data were originally generated are shown in black in Figures 3C,D.

The solutions shown in Figure 3 are among the infinitely many admissible solutions generated by the following procedure, which actually only yields a small subset of all possible solutions. Starting at some initial point in the Gamma-space, a phase-plane trajectory is computed according to a stable linear state-space model of the form $\dot{\boldsymbol{\gamma}}(t) = B\boldsymbol{\gamma}(t)$. This is certainly not the only strategy for creating flux sets, but it constitutes a simple option that leads to continuous fluxes. A Monte-Carlo approach is utilized, in which a 2×2 matrix B is randomly generated, but where only those matrices B are retained that have negative real eigen values and result in non-negative fluxes for all time points. The resulting set of trajectories yields many dynamical fluxes with quite different features. Figure 3C shows some feasible solutions for fluxes v_1 through v_4 in multiple colors as thin lines, superimposed on the flux of the actual model, from which the concentration data were generated (black line). These fluxes are shifted in Panel (C), so that their initial values match, in order to facilitate easier comparisons. Interestingly, the resulting fluxes can possess behaviors ranging

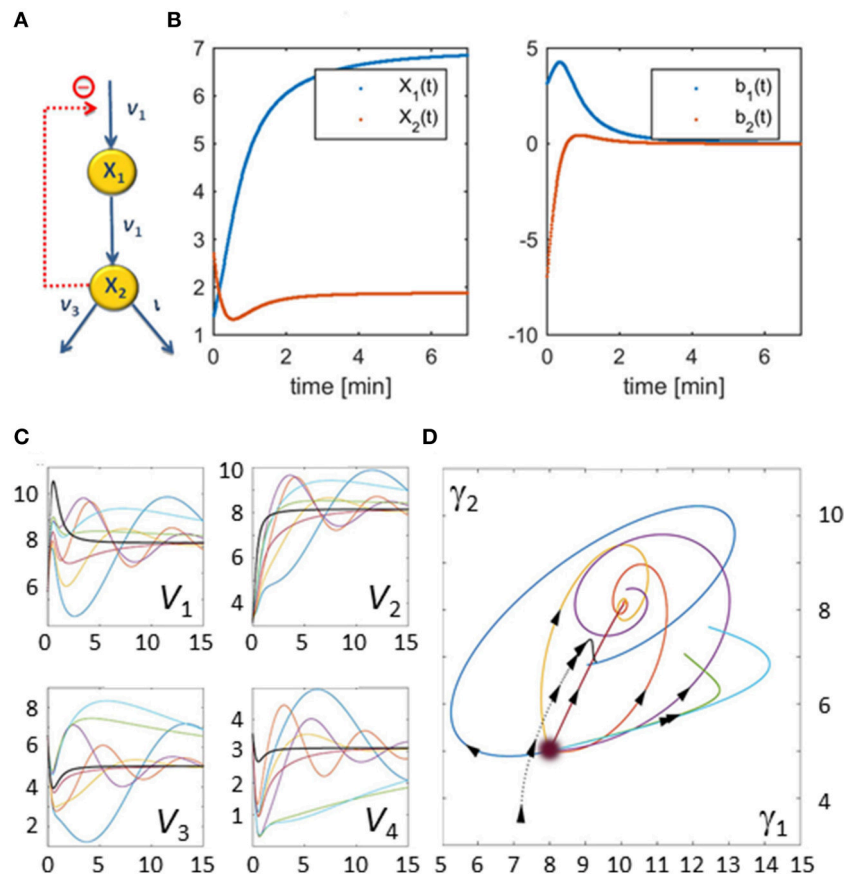


FIGURE 3 | Illustration example used to demonstrate the core concepts of the flux characterization procedure. The pathway has a very simple structure as depicted in Panel (A). Panel (B) shows $X_1(t)$ and $X_2(t)$ on the left and the slopes of $X_1(t)$ and $X_2(t)$ estimated from noise-free measurements on the right. Panel (C) shows seven examples of flux sets vs. time that satisfy Equation (5) exactly; for this illustration, all start at the same point as the original flux set ($\gamma(0) = [6.3271, 3.1588, 6.5486, 3.5486]$ corresponding to $\gamma(0)^T = [8, 5]$). The thicker black curves are the fluxes with which the original data were produced. The corresponding Gamma-trajectories are depicted with the same color scheme in Panel (D). The blurry dot indicates the common start value of these trajectories while the dotted line represents the true flux, which is known in this artificial example.

from simple shoulder curves to over- and undershoots and different oscillatory responses. One notes that this Monte-Carlo strategy does not address issues of noise in the data, but is simply a means of retrieving diverse solutions that are mathematically admissible.

Admissible Subset of Gamma-Space: The Subspace of Non-Negative Fluxes

For biological realism, it is necessary to determine the set of γ 's for which the corresponding vector $v(t)$ consists of non-negative values for all fluxes and all time points. According to Equation (4), the feasible space, given by $v(t) = A^+b(t) + \text{null}(A)$ $\gamma > 0$, is an intersection of m half-spaces:

$$A^+(i, :)b(t) + \gamma_1 \text{vec}_{1,i} + \dots + \gamma_d \text{vec}_{d,i} \geq 0 \quad i = 1, 2, \dots, m \quad (6)$$

Here, $A^+(i, :)$ denotes the i^{th} row of the $m \times n$ Moore-Penrose pseudo-inverse matrix. The inequalities are linear and thus constitute a bounded or unbounded polytope.

Formulating the Problem as an Optimization Task

According to Equation (6), the solution set is still infinite, thus raising the question of whether biological constraints could be evoked to reduce the feasible space of solutions. A possibly pertinent constraint for the selection of meaningful flux profiles is the overall minimization of the magnitudes of positive fluxes, which might be interpreted as a form of metabolic energy conservation. Minimizing the sum of fluxes at steady state has been referred to as the *parsimonious enzyme effect* (Lewis, 2010). Here, the terminology is slightly different, as the minimization is done for the sum of all fluxes over all time points. Since the non-negativity constraints are already in place, this sum of fluxes at all time points equals the so-called “minimum L_1 ” or “Manhattan-” norm, which is defined as $\min_{\substack{\gamma \geq 0 \\ Av=b}} \|v\|_1 = \min_{\substack{\gamma \geq 0 \\ Av=b}} \sum_{i=1}^m |v_i| =$

$\min_{\substack{\gamma \geq 0 \\ Av=b}} \sum_{i=1}^m v_i$. The optimization problem leading to this result in terms of γ is shown in Equation (7). The constraint $Av = b$ is

already taken into account, since the representation in Equation (4) only allows for fluxes that satisfy this constraint. Thus, the optimization simplifies to:

$$\begin{aligned} \min_{A^+b(t)+\text{null}(A)\gamma(t)\geq 0} \sum_{i=1}^m A^+b(t) + \text{null}(A)\gamma(t) \\ = \min_{A^+b(t)+\text{null}(A)\gamma(t)\geq 0} \sum_{i=1}^m \text{null}(A)\gamma(t) \end{aligned} \quad (7)$$

The important insight from Equation (7) is that the optimization problem can be translated into a simpler linear program in terms of $\gamma(t)$, which can be solved using algorithms for linear programming, such as the simplex method. In practice, testing the corner points of the feasible polyhedron for identifying the corner with the minimum sum is a very well-established way of arriving at the optimal solution (Dantzig, 1984).

One should note that DFE and the choice of an objective function for the identification of biologically reasonable flux solutions are entirely independent. For instance, as an alternative optimization approach to minimizing the sum of fluxes for all time points, we could select the L_2 -norm of the flux vector at each point in time. This choice emphasizes and weighs the roles of the individual fluxes in a different manner. Minimizing the squared sum of fluxes at steady state has been referred to as *flux optimization* (Holzhütter, 2004). Again, our terminology is slightly different because the minimization pertains to all fluxes and all time points. This task is described in Equation (8) and again represents in some sense the minimum-energy flux set.

$$\begin{aligned} \min \quad & \|\gamma\|_2^2 \\ \text{s.t.} \quad & \gamma \geq 0 \\ & A\gamma = b \end{aligned} \quad (8)$$

The optimization problem in Equation (8) can be reformulated as the optimization problem of minimizing the L_2 -norm of the vector $\gamma(t)$. Equation (9) shows this reformulation:

$$\begin{aligned} \min_{A^+b(t)+\text{null}(A)\gamma(t)\geq 0} (A^+b(t) + \text{null}(A)\gamma(t))^T (A^+b(t) \\ + \text{null}(A)\gamma(t)) = \min_{A^+b(t)+\text{null}(A)\gamma(t)\geq 0} (A^+b(t))^T A^+b(t) \\ + \gamma(t)^T \text{null}(A)^T \text{null}(A)\gamma(t) + \gamma(t)^T \text{null}(A)^T A^+b(t) \\ + (A^+b(t))^T \text{null}(A)\gamma(t) = \min_{A^+b(t)+\text{null}(A)\gamma(t)\geq 0} \gamma(t)^T I_m \gamma(t) \\ = \min_{A^+b(t)+\text{null}(A)\gamma(t)\geq 0} \|\gamma(t)\|_2^2 \end{aligned} \quad (9)$$

Here, $\text{null}(A)^T \text{null}(A) = I_m$ is the identity matrix of dimension m , because the columns of $\text{null}(A)$ are orthonormal base vectors of the null space. Furthermore, the pseudo-inverse solution $A^+b(t)$ is orthogonal to the null space, so that $\text{null}(A)^T A^+b(t) = (A^+b(t))^T \text{null}(A) = 0$. Additionally, $(A^+b(t))^T A^+b(t)$ does not change with $\gamma(t)$, so that its removal from the optimization problem does not change the result. Thus, it is of note that Equation (9) is equivalent to the quadratic program of Equation (8).

Other optimization problems could be formulated, but the interesting challenge is that it is not really known what “optimality” means for the fluxes in a biological system or organism. Optimal solutions, with respect to various criteria, could be suggested, but whether these solutions are compatible with additional information about the functional form or about effectors of fluxes needs to be tested for specific problems. A later section examines the minimum-energy solution for a realistic biological system and indeed challenges the validity of this particular solution to some degree. This discussion shows that optimization, which at this stage does not assume any functional form for the fluxes, may lead to fluxes that can be questionable. At the same time, these optimal solutions can be utilized as starting points for approaching solutions that appear to be biologically meaningful.

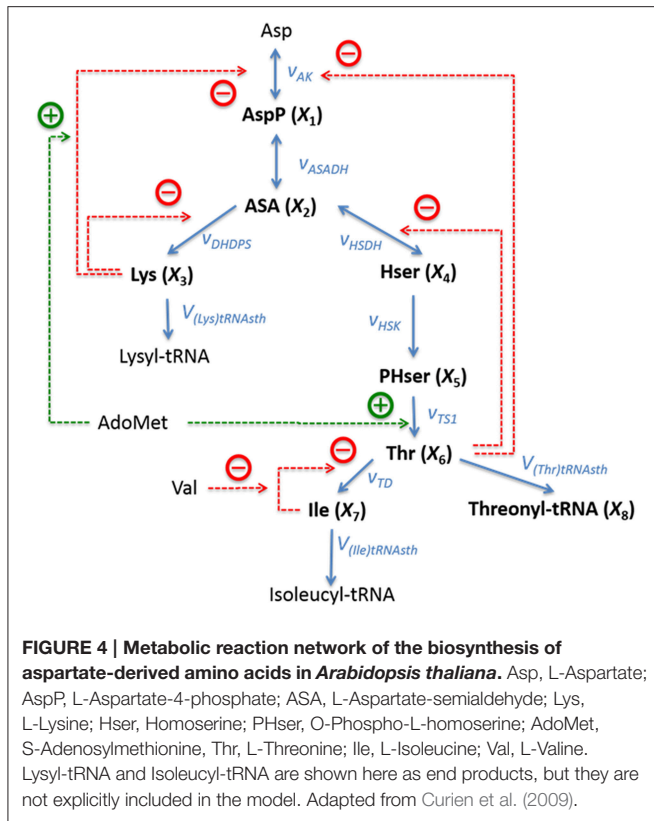
Illustration Example: The Biosynthetic Pathway of Aspartate-Derived Amino Acids in the Plant *Arabidopsis thaliana*

After characterizing a feasible set of fluxes, optimizing the parameters for these fluxes yields a reasonable default solution. Nonetheless, accounting additionally for generally expected features of fluxes can lead to more biologically relevant flux sets. Such generic features may include knowing that a certain flux is a function of only one variable, i.e., its substrate. Another piece of generic information could be that, when a substrate of a flux is zero, the flux has to equal zero as well. These types of constraints are illustrated below with a specific example from the literature, namely the biosynthetic pathway of aspartate-derived amino acids in the plant *Arabidopsis thaliana* (Curien et al., 2009). In reference to the lead author of a model of this system, we will call it the “Curien” model. Since the complete model and the fluxes are known, the pathway system constitutes a good test case. The Gamma-trajectory for the Curien model will be plotted, the criterion of non-negativity and its implication in Gamma-space will be investigated and determined, and the result of optimization will be studied and compared to the original fluxes. Finally, auxiliary methods of flux improvement will be suggested.

Identification of Flux Trends

The pathway of biosynthesis of aspartate-derived amino acids is responsible for the distribution of the carbon influx into the synthesis of threonine, lysine, methionine, and isoleucine (Figure 4). The original kinetic model (Curien et al., 2009) was constructed based on *in vitro* kinetic measurements, assuming generalized functional forms of the fluxes in the tradition of Michaelis and Menten. The model contains seven dependent variables, namely, X_1 = [aspartyl-phosphate], X_2 = [aspartate semialdehyde], X_3 = [lysine], X_4 = [homoserine], X_5 = [phosphohomoserine], X_6 = [threonine], and X_7 = [isoleucine]. Additionally we consider the output variable X_8 = [threonyl-tRNA].

This specific example is well-suited as an illustration of the proposed techniques of flux identification, because it is representative and of moderate complexity, and because its



details are fully known, which facilitates method development and multiple diagnoses of problems that are likely to arise.

The equations for the model

$$\begin{aligned}
 \frac{dX_1}{dt} &= v_{AK} - v_{ASADH} \\
 \frac{dX_2}{dt} &= v_{ASADH} - v_{DHDPS} - v_{HSDH} \\
 \frac{dX_3}{dt} &= v_{DHDPS} - v_{(Lys)tRNAsth} \\
 \frac{dX_4}{dt} &= v_{HSDH} - v_{HSK} \\
 \frac{dX_5}{dt} &= v_{HSK} - v_{TS1} \\
 \frac{dX_6}{dt} &= v_{TS1} - v_{TD} - v_{(Thr)tRNAsth} \\
 \frac{dX_7}{dt} &= v_{TD} - v_{(Ile)tRNAsth} \\
 \frac{dX_8}{dt} &= v_{(Thr)tRNAsth}
 \end{aligned} \quad (10)$$

are directly taken from the original article. The functional forms of the fluxes are presented in Equation (11):

$$v_{AK1} = [AK1] \cdot \frac{5.65 - 1.6[AspP]}{1 + ([Lys]/(\frac{550}{1+[AdoMet]/3.5}))^2}$$

$$\begin{aligned}
 v_{AK2} &= [AK2] \cdot \frac{3.15 - 0.86[AspP]}{1 + ([Lys]/22)^{1.1}} \\
 v_{AKI} &= [AKI - HSDH I] \cdot \frac{0.36 - 0.15[AspP]}{1 + ([Thr]/124)^{2.6}} \\
 v_{AKII} &= [AKII - HSDH II] \cdot \frac{1.35 - 0.22[AspP]}{1 + ([Thr]/109)^2} \\
 v_{AK} &= v_{AK1} + v_{AK2} + v_{AKI} + v_{AKII} \\
 v_{ASADH} &= [ASADH] \cdot (0.9[AspP] - 0.23[ASA]) \\
 v_{HSDH I} &= [AKI - HSDH I] \cdot 0.84 \cdot \left(0.14 + \frac{0.86}{1 + [Thr]/400}\right) \\
 v_{HSDH II} &= [AKII - HSDH II] \cdot 0.64 \cdot \left(0.25 + \frac{0.75}{1 + [Thr]/8500}\right) \\
 v_{HSDH} &= v_{HSDH I} + v_{HSDH II} \\
 v_{DHDPS1} &= [DHDPS1] \cdot [ASA] \cdot \frac{1}{1 + ([Lys]/10)^2} \\
 v_{DHDPS2} &= [DHDPS2] [ASA] \cdot \frac{1}{1 + ([Lys]/33)^2} \quad (11) \\
 v_{DHDPS} &= v_{DHDPS1} + v_{DHDPS2} \\
 v_{(Lys)tRNAsth} &= V^{AaRS} \cdot \frac{[Lys]}{25 + [Lys]} \\
 v_{HSK} &= [HSK] \cdot \frac{2.8[Hser]}{14 + [Hser]} \\
 v_{TS1} &= [TS1] \cdot \left(\frac{0.42 + 3.5[AdoMet]^2/73}{1 + [AdoMet]^2/73}\right) [PHser] \\
 v_{(Thr)tRNAsth} &= V^{AaRS} \cdot \frac{[Thr]}{100 + [Thr]} \\
 v_{TD} &= [TD] \cdot \frac{0.0124[Thr]}{1 + ([Ile]/(30 + \frac{74[Val]}{610 + [Val]}))^3} \\
 v_{(Ile)tRNAsth} &= V^{AaRS} \cdot \frac{[Ile]}{20 + [Ile]}
 \end{aligned}$$

Equation (10) can equivalently be written in vector form as shown in Equation (7), namely as

$$\frac{dX}{dt} = \dot{X} = A v \quad (12)$$

where v and A are the corresponding vector of reaction rates (i.e., fluxes) and the stoichiometric matrix, respectively. For the Curien model, they are shown in Equations (13) and (14):

$$\begin{aligned}
 v &= [v_{AK}, v_{ASADH}, v_{HSDH}, v_{DHDPS}, v_{(Lys)tRNAsth}, v_{HSK}, v_{TS1}, \\
 &\quad v_{(Thr)tRNAsth}, v_{TD}, v_{(Ile)tRNAsth}]^T \\
 &= \{v_1, v_2, v_3, v_4, v_5, v_6, v_7, v_8, v_9, v_{10}\}^T \quad (13)
 \end{aligned}$$

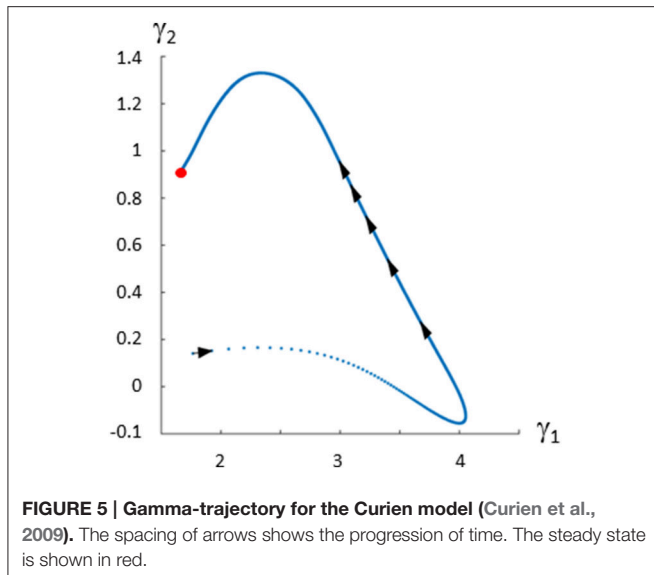


FIGURE 5 | Gamma-trajectory for the Curien model (Curien et al., 2009). The spacing of arrows shows the progression of time. The steady state is shown in red.

$$A = \begin{pmatrix} 1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & -1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & -1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & -1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & -1 & -1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & -1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \end{pmatrix} \quad (14)$$

Gamma-Trajectory of the Curien Model

The fluxes and metabolite concentrations for this system are known, which allows us to plot the “true” Gamma-trajectory in the Gamma-space representation vs. time:

$$\mathbf{v}(t) = A^+ \mathbf{b}(t) + \text{null}(A) \boldsymbol{\gamma} \quad (15)$$

Here,

$$\begin{aligned} \text{null}(A) &= [\text{vec}_1, \text{vec}_2] \\ &= \begin{bmatrix} 0.5374 & 0.5374 & 0.1162 & 0.4212 & 0.4212 & 0.1162 & 0.1162 & 0 & 0.1162 & 0.1162 \\ 0.0534 & 0.0534 & 0.3914 & -0.3380 & -0.3380 & 0.3914 & 0.3914 & 0 & 0.3914 & 0.3914 \end{bmatrix}^T \end{aligned}$$

spans the null space of A . This solution is easily found, as $\text{null}(A)$ is a MATLAB command that returns these two orthonormal vectors. $\boldsymbol{\gamma}(t) = [\gamma_1(t), \gamma_2(t)]^T$ is the vector of coefficients associated with $\text{null}(A)$. With this information, the two-dimensional Gamma-space can be explored instead of the feasible subset of the 10-dimensional space of fluxes.

For each time point t , the gamma coefficients can be calculated by projecting $\mathbf{v}_{\text{null}}(t) = \mathbf{v}(t) - A^+ \mathbf{b}(t)$ onto the vectors vec_1 and vec_2 . The result is equivalent to the dot product of $\text{null}(A)$ and $\mathbf{v}(t)$, since $A^+ \mathbf{b}(t)$ is orthogonal to the null space and the dot product is zero.

Figure 5 shows the trajectory starting at time zero and ending at steady state shown with a red dot.

Feasible Solutions

Similar to the introductory example, this model permits an infinite number of solutions, which may be quite different. Some of these feasible solutions can be generated with a Monte-Carlo simulation by starting at some initial point in the Gamma-space and computing a phase-plane trajectory according to the linear state-space model of $\dot{\boldsymbol{\gamma}}(t) = B\boldsymbol{\gamma}(t)$, as before. The resulting trajectories exhibit a variety of different dynamical characteristics for the fluxes. Panels 1–9 of **Figure 6** show in multiple colors a selection of feasible solutions for fluxes v_1 through v_{10} , with the exception of the output flow v_8 . Flux v_8 is not shown since it belongs to the only full rank subset of the system and is fully determined by numerically differentiating X_8 . The thin lines representing these solutions are superimposed on the actual flux (black), which is known from the model. It is evident that some of the inferred fluxes are similar to the actual fluxes, but that many are not even qualitatively of the same shape. In order to facilitate easier comparisons, the fluxes shown are shifted so that their initial values match. Interestingly, the inferred fluxes show different behaviors ranging from monotonic to various oscillatory shapes. One should note that these feasible solutions are typical examples if we assume a trajectory from a linear state-space solution but that they by no means represent all the possible trends.

An interesting observation is that one may add an equal value to each flux in *Set 1* = $\{v_1, v_2, v_4, v_5\}$ and/or *Set 2* = $\{v_1, v_2, v_3, v_6, v_7, v_9, v_{10}\}$ without a change in the metabolite concentration profiles. The reason is that these shifts cancel out in the original differential equations (Equation 10) and $\dot{\mathbf{X}}(t)$ therefore stays the same. **Figure 7A** demonstrates that the shape of the Gamma-trajectory (**Figure 5**) can be shifted along the red line if one adds different positive constant amounts to *Set 1* and along the cyan line if one adds different positive constant amounts to *Set 2*. Of course, shifts in both directions are admissible as well. One could also pick negative constant values as long as the fluxes stay positive. This way, the entire Gamma-space can be spanned. This is an equivalent, and perhaps more comprehensible, explanation of the two degrees of freedom for this pathway. As an alternative to constant shifts, it is even

admissible to add the same function of time to all fluxes in the sets.

Admissible Subset of the Gamma-Space: the Subspace of Non-Negative Fluxes

For each time point t , we determine the set of $\boldsymbol{\gamma}$'s for which the corresponding $\mathbf{v}(t)$ consists entirely of non-negative fluxes. Recalling Equation (6), the feasible space here is an intersection of 10 half spaces characterized by the following set of inequalities:

$$A^+(i, :) \mathbf{b}(t) + \gamma_1 \text{vec}_{1,i} + \gamma_2 \text{vec}_{2,i} \geq 0 \quad i = 1, 2, \dots, 10 \quad (16)$$

Here, $A^+(i, :)$ denotes the i^{th} row of the 10×8 Moore-Penrose pseudo-inverse matrix.

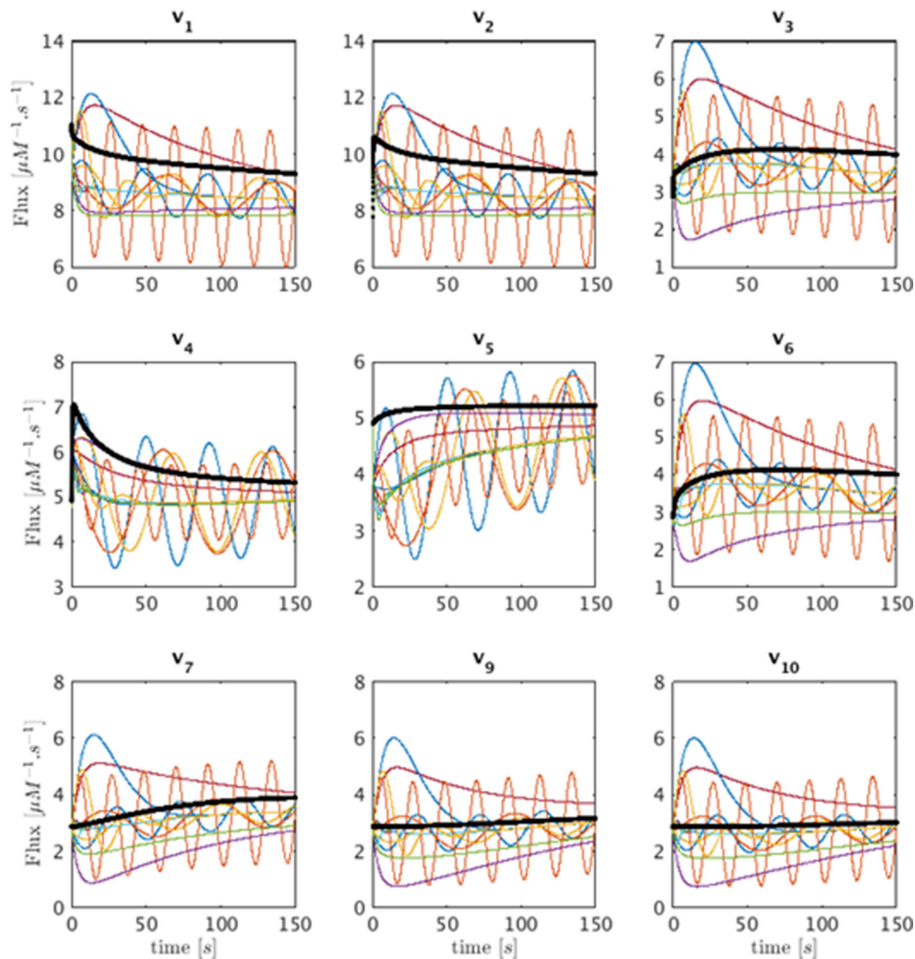


FIGURE 6 | Sets of feasible solutions for each flux v_1 to v_7 and v_9 to v_{10} is shown in each panel. For visualization purposes, the actual flux from the model is shifted to have the same initial value as the simulated fluxes and is superimposed as a thick black line for comparison.

In this example, only two out of the total of 10 inequalities happen to be active inequalities, which results in a feasible subspace in the shape of an open triangle. One should note, however, that $b(t)$ changes with time, so that there is a new open triangle for each time point. Expressed differently, the feasible region resulting in non-negative flux sets varies with each time point. **Figure 7B** exhibits the first seven of these open triangles in different shades of red. There is one such triangle for each time point; the triangles are not shown for the following time points to avoid over-population of the plot.

The corners of these open triangles are shown as black dots, which lie on a curve. The blue curve shows the actual Gamma-trajectory of **Figure 5**. One interesting observation is that, for the initial time points, the two curves (“true” and inferred) coincide. For later time points, the blue curves lie inside the corresponding open triangle of non-negative solutions.

Any continuous trajectory whose points fall inside these non-negative open triangles for all time points is a feasible flux profile.

Minimum-Energy Flux Set

Searching the feasible solutions for the set of flux profiles that minimize the sum of squared flux norms for all time points results in the minimum-energy flux. This procedure is equivalent to solving the quadratic programming of Equation (9) and results in the same flux profile as solving the linear programming of Equation (7). For the case of the Curien model, both of these methods yield the same set of fluxes as the corner solution introduced in the previous section. This solution is also equivalent to the result of a non-negative least-squares optimization problem performed in MATLAB.

Figure 8 shows the minimum energy flux profiles plotted vs. time (depicted in red) together with the actual fluxes of the Curien model (blue). The two solutions are quite different, although they both match the metabolite data perfectly. The next sections introduce strategies to alleviate this discrepancy. One should note that the computed solution is actually “cheaper” than the Curien model, as all fluxes have lower magnitudes; whether it is “better” or “worse” than the Curien model cannot be said, because we do not know the correct criteria.

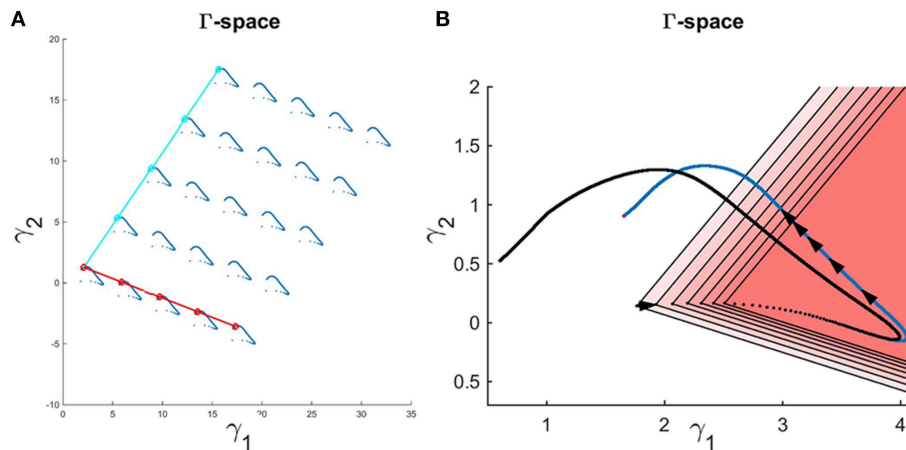


FIGURE 7 | (A) Adding a constant amount to the fluxes in Set 1 for all time points shifts the Gamma-trajectory along the dark red line without any change in the concentration profiles for all metabolites. Similarly, adding a constant amount to the fluxes in Set 2 for all time points shift the Gamma-trajectory along the cyan line without any change in the concentration profiles for all metabolites. **(B)** The Gamma-trajectory of the Curien model is depicted in blue color. The black arrowheads shown halfway through the blue curve are equally spaced in time. The open red triangles show the subset of the Gamma-space where the corresponding flux set is non-negative at each point in time. Only the first seven triangles are shown for illustration purposes. The black dotted curve shows the corners of these open triangles for different time points. We will later see that, for the Curien model example, this curve is the same as the minimum-energy curve as described before. Interestingly the blue and black curves are overlapping in the beginning but then diverge.

Generally Expected Features Regarding Fluxes Can Restrict the Feasible Space Further

General expectations regarding metabolic fluxes may constrain the feasible flux profiles. To assess these expectations, it is useful to plot the fluxes against their substrates and modulators rather than against time, as was done before. **Figure 9** shows all actual fluxes plotted against their substrates and effectors in blue, super-imposed on the min-energy fluxes vs. their substrates and effectors in red. Fluxes v_5 , v_6 , v_7 , and v_{10} are known to be functions solely of their corresponding substrates, while fluxes v_2 , v_3 , v_4 , and v_9 have two substrates/regulators, and v_1 has three. Closer inspection of these plots reveals that the plots of v_6 vs. X_4 and v_7 vs. X_5 show a behavior that is not consistent with a true mathematical function, namely a folding-over (**Figure 9A**). For example, if the concentration of X_4 is 1.2 μM , flux v_6 may take two values, and therefore cannot be a function in the mathematical sense. Assuming that we know that no other variables affect this flux, this folding-over phenomenon is not acceptable.

To ameliorate this problem, one may remove or cut the folded-over section. Specifically, for the time points corresponding to folded-over values, we let v_6 take values according to the top branch. This is allowable, as the upper branch is a feasible solution. Using this technique, $v_6(t)$ becomes uniquely determined and can be considered an identified flux. Subsequently, a new min-energy response can be computed with exactly the same methods as before, but with only one degree of freedom left.

Figure 10 depicts the same plots as in **Figure 9** after removing the folding-over phenomenon. Interestingly, all fluxes in Set 2 = $\{v_1, v_2, v_3, v_6, v_7, v_9, v_{10}\}$, as introduced before, are now fixed and almost equivalent to the actual fluxes. This means that the number of degrees of freedom has decreased to 1 after

incorporating the information that one of the fluxes is a function of one variable only. The discrepancy between fluxes in Set 1 = $\{v_1, v_2, v_4, v_5\}$ remains unsolved, and there is no other folding-over among the one-variable fluxes.

A caveat of the strategic step above is our assumption that some of the fluxes only depend on their substrates. Such an assumption is of course not always valid, but the more we learn about metabolism the more we will be able to rely on solid information. To validate such an assumption, one might use a step-wise scheme of testing additional variables as modulators (Marino and Voit, 2006). By the same token, the proposed methods may actually point to regulatory signals that had been unknown or overlooked (Dolatshahi et al., 2016a,b). One notes that this issue is a challenge for any estimation or identification strategy.

In order to recover the fluxes in Set 1, additional information is needed. First, one could assume that all fluxes in this set are shifted by the same value. If this value were chosen as about 0.3, one can imagine from **Figure 11** that the fluxes become very similar to the fluxes in the original model. Second, suppose it was known that, for instance, v_5 is well-modeled as a Michaelis–Menten rate function and the corresponding kinetic parameters K_M and V_{max} could be extracted from the literature. Then one could find v_1 , v_2 , v_4 by the following simple procedure: Determine the shift function $f_{\text{shift}}(t) = \frac{V_{\text{max}} X_3(t)}{K_M + X_3(t)} - v_{5-\text{min}}(t)$ and add it to the rest of fluxes in Set 1 to find the actual fluxes; thus, $v_j(t) = v_{j-\text{min}}(t) + f_{\text{shift}}(t)$, $j \in \{1, 2, 4\}$. Indeed, if the Michaelis–Menten function is implemented with Curien's parameter values, the entire system is perfectly recouped (result not shown). Having said that, there is no objective argument against the fluxes in **Figure 11**, except possibly that v_5 is essentially 0 for the first 250 time units, and then becomes

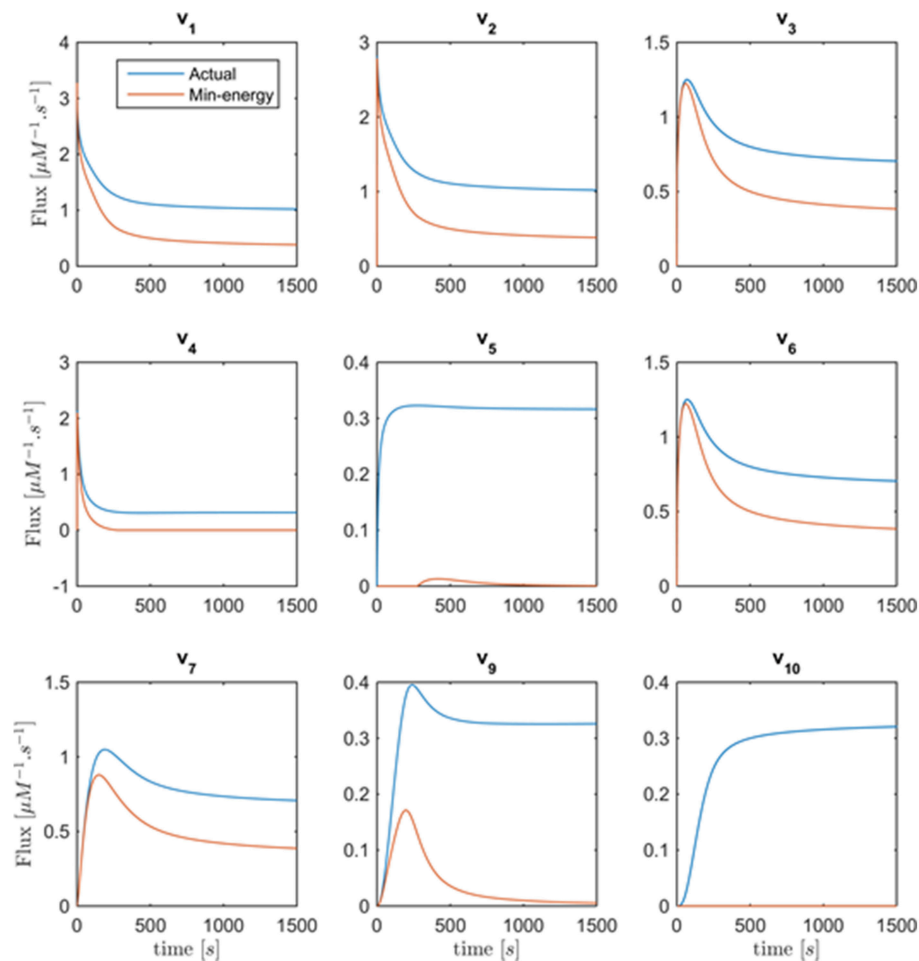


FIGURE 8 | Fluxes v_1 to v_{10} with the exception of v_8 are plotted vs. time. Curves in red are the min-energy fluxes, while the blue curves show the actual fluxes of the Curien model. Flux v_8 is not shown because it belongs to the full-rank subset of the system and can be recovered exactly from the data.

slightly non-monotonic, which might not be realistic. At the same time, the computed fluxes are of lower magnitude than those in the Curien model. As a third alternative, one could independently determine one of the fluxes in *Set 1*, for example as a power-law function, as it was demonstrated elsewhere (Iwata et al., 2013), and then compute all other fluxes of the set.

DISCUSSION

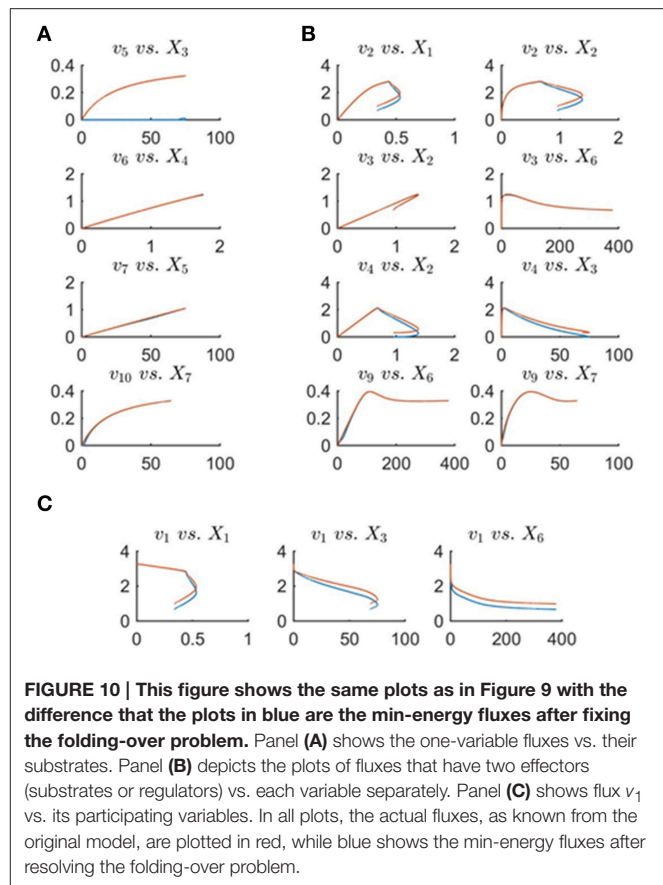
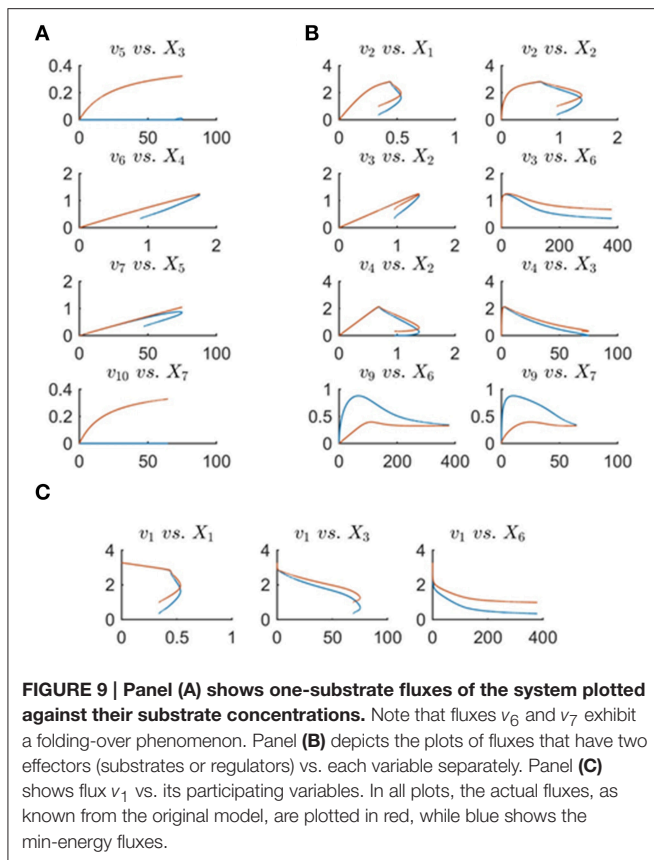
Extension of DFE Toward Pathways with Incomplete Information

In many practical scenarios, some of the data are missing, and/or some of the fluxes cannot be determined fully even with the techniques described in the previous sections. If so, the need arises for additional strategies that make maximal use of DFE's capabilities and diagnostic features (Voit, 2009; Chou and Voit, 2012; Iwata et al., 2013), along with random search and global optimization techniques.

Because data are seldom ideal, this section discusses a rather generic, multi-step strategy that takes advantage of the diagnostic and computational benefits that DFE offers, and augments them with auxiliary methods and global optimization approaches toward a full-system parameterizations (Figure 12). These procedures were recently used for the construction of a complex model of the highly regulated glycolytic pathway of *Lactococcus lactis* from NMR data (Dolatshahi et al., 2016a,b) where, due to missing data and other features of the data, the estimation of parameters was not straightforward.

The first step of this strategy consists of identifying full rank subsets of fluxes within the system (see flux estimation module in Figure 12), if that is possible. For instance, the *Arabidopsis* example allowed us to identify *Set 2* as well as the flux v_8 .

Suppose now that data for one or more of the variables are missing. If so, the “missing metabolite estimation module” in Figure 12 is used (see also Voit, 2009). The goal is to infer flux information from near-by metabolites or at least to constrain the parameters of this flux for the following steps of a randomized



search and global full system parameterization. This module involves an optimization task, which ideally yields valuable information regarding the likely profile of the missing data. The first step in this module consists of selecting a metabolite pool that is close to the missing data, includes a concentration profile, and has influxes and effluxes that are at least partially characterized. As an example, assume time series data for lysine (X_3) were missing in the Curien model. The idea is to infer the missing data from other metabolites and/or identifiable fluxes. For instance, information regarding lysyl-tRNA could provide valuable hints regarding $V_{(Lys)tRNAsth}$; namely, one could assume a power-law or Michaelis–Menten function to infer X_3 from the data for the accumulation of lysyl-tRNA. In this particular case, the computation of X_3 from $V_{(Lys)tRNAsth}$ at different time points would actually be quite simple, as both functional formats can be transformed into linear equations.

If such an inference is not feasible, other biological information is needed and must be supplied on a case-by-case basis. For instance, biological arguments may provide clues regarding amounts that might reasonably be added to formerly identified flux sets. In some cases, measurements fall below the detection limit, so that no numerical data are available, although the biology of the system mandates that the concentrations are not zero. The detection limit, mass conservation, and possibly other considerations can serve as useful constraints for the optimization algorithm. The output of this module thus consists of substitutes for some of the missing

data profiles, along with their associated parameter values. In other parts of the workflow, these are treated like experimental data.

The “validation of functional form and regulation” step assesses the appropriateness of the functional formats for the flux representations. A first and obvious criterion is the quality of the fit, which is necessary, although not sufficient (Voit, 2011). A second criterion is the detection or lack of “runs in residuals” (Draper and Smith, 1981). If no appropriate format and parameterization can be found, it is quite probable that important components of the pathway are missing from the model. An example is the situation where a flux decreases with increasing, reasonable substrate concentrations. Such a trend is counterintuitive and may suggest that a regulator is missing from the model. If so, DFE can possibly help identify what shape the dynamic trend of the regulator must have to remedy the discrepancy. A scan of the dynamics of all variables in the model may even identify candidates, although such inferences are still to be tested experimentally. Examples of this situation are presented elsewhere (Dolatshahi et al., 2016a,b).

Beyond the quality of fit and run test, no true validation is possible, because the fluxes are unknown. Even so, the “validation of functional form and regulation” step ensures reasonableness and flags fluxes that are computed as negative, exhibit unduly high magnitudes, or are apparently lacking important contributing variables.

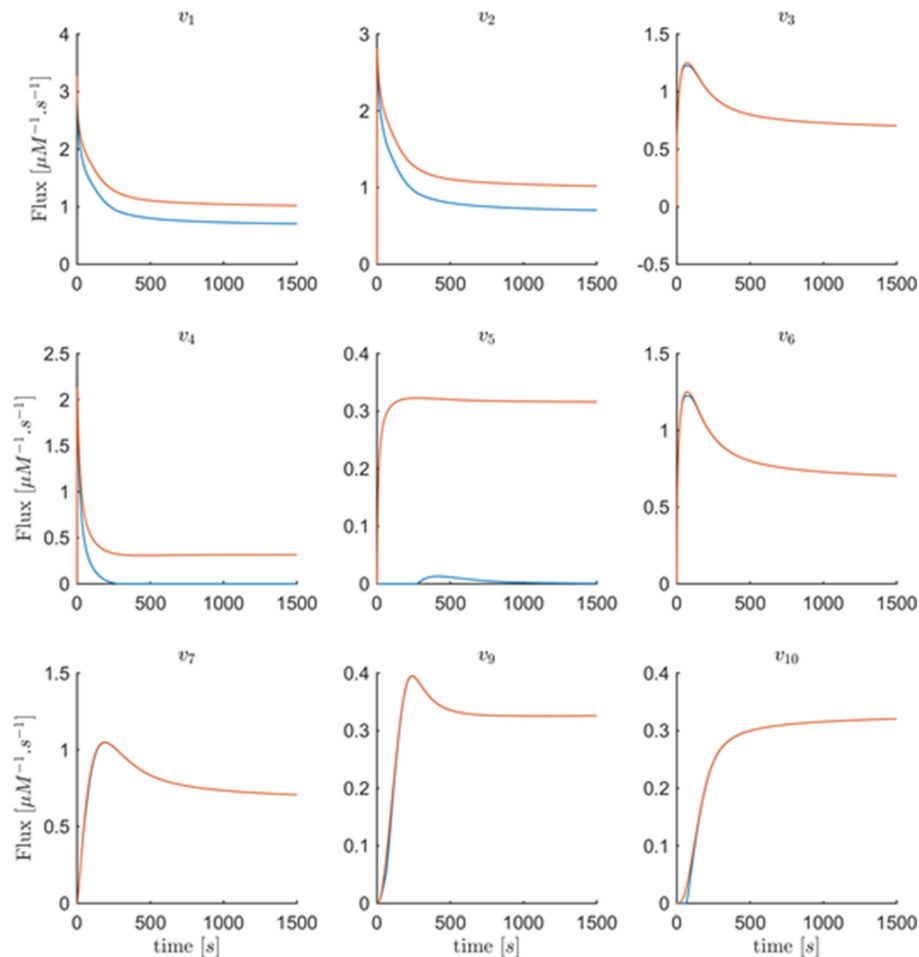


FIGURE 11 | Fluxes v_1 to v_{10} with the exception of v_8 are plotted vs. time. The red curves are the min-energy fluxes after solving the folding-over problem, while the blue curves show the actual fluxes. It is evident that the fluxes v_3 , v_6 , v_7 , v_9 , v_{10} are almost identical and overlapping and that our method has recovered these fluxes.

Assessment of the Inferred Fluxes and their Parameters

Once the functional forms and regulations are considered satisfactory and the corresponding parameters are estimated, it is necessary to test whether the estimated parameter set is essentially unique or whether substantially different solutions exist. This identifiability and sloppiness step (e.g., Gutenkunst, 2007a,b; Vilela, 2009; Raue, 2013; Villaverde and Banga, 2013; Tafintseva, 2014; Tönsing et al., 2014) is particularly pertinent if the data are noisy or some of the data were not measured but inferred in earlier steps. This global analysis often utilizes Monte Carlo simulations, in which a large-scale random search is anchored in the estimated, optimal parameter set $\{P_i\}$, which serves as the starting point for the global optimization. The differences in the sets of newly estimated parameter values for each flux and each experiment are collectively used to determine admissible ranges for the parameters of the system and starting values for global optimization. This last estimation step entails a combination of different optimization techniques,

which may begin with evolutionary (genetic) algorithms that provide coarse solutions and are followed up with steepest descent algorithms that refine these solutions. The objective function for this purpose is the usual sum of squared errors over all time points, metabolites, and datasets, but may also include a penalty for metabolite concentrations that were inferred rather than directly measured. The ideal outcome of this step is either an essentially unique model parameterization or a compact ensemble of models with parameter values that permit some flexibility without compromising the data fit.

CONCLUSIONS AND OUTLOOK

The goal of this article was to extend the utility of DFE to the relatively common scenario where the algebraic system of fluxes is underdetermined or some time series data are missing or incomplete.

Initially, the concept of lower-dimensional representation in the form of a so-called Gamma-space and a Gamma-trajectory

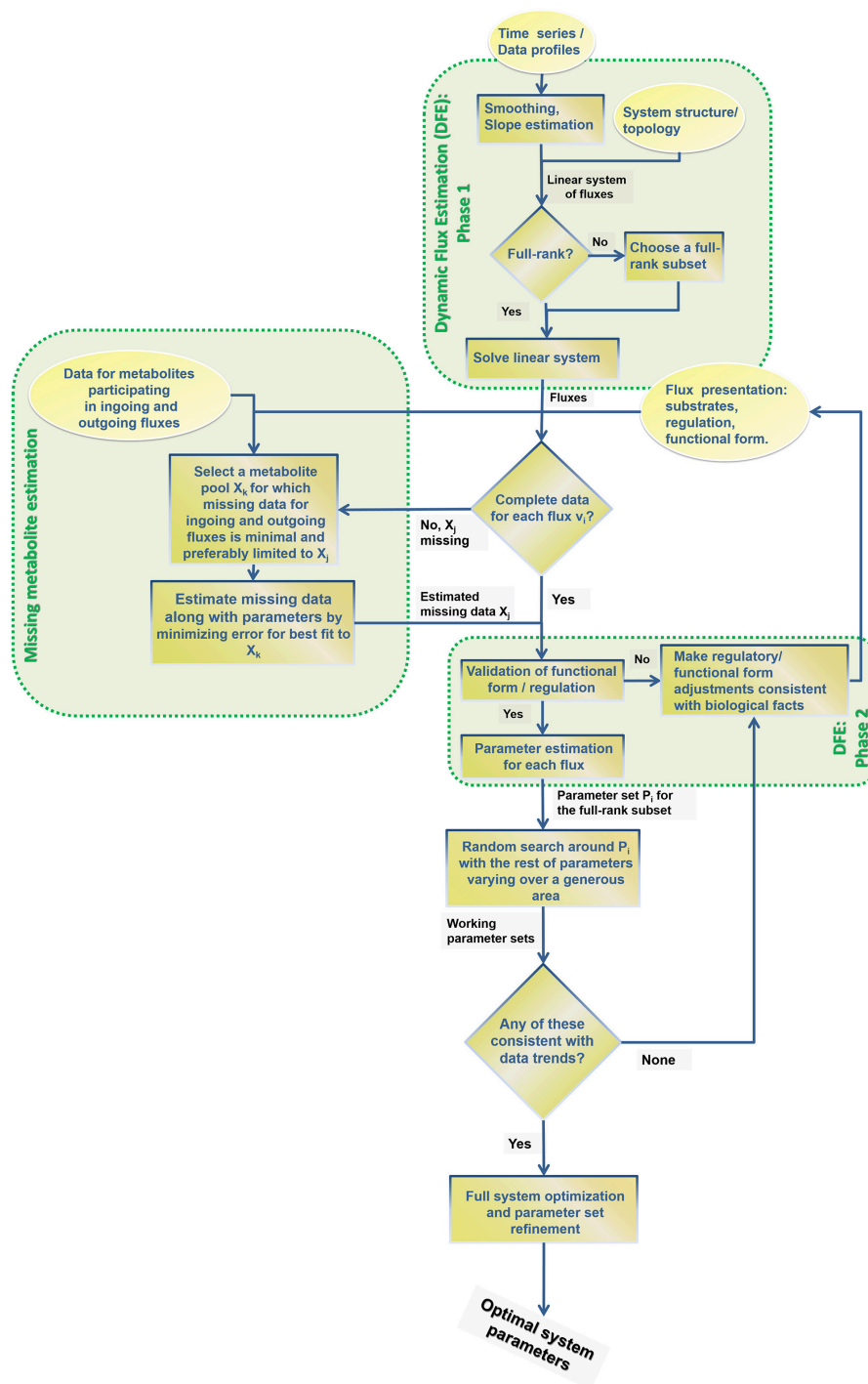


FIGURE 12 | Step-by-step procedure for the proposed extension of dynamic flux estimation (DFE).

was introduced. This representation is especially useful when the number of degrees of freedom is low. Reasonable biological constraints like smoothness over time and non-negativity of fluxes were taken into account to constrain the feasible space even further. In particular, a minimum-energy criterion was

considered, and solutions were discarded in which fluxes were not representable by mathematical functions, due to non-uniqueness. The concepts were illustrated with a model of aspartate metabolism in the plant *Arabidopsis*. The minimum-energy flux set did not match the actual flux profiles for this

pathway, even though the metabolite data were recouped with a set of fluxes that had lower magnitudes than in the original model. The addition of biologically reasonable constraints reduced the discrepancies. In particular, it was known that a certain flux, v_6 , is a function of only its substrate. This knowledge helped us reshape the minimum-energy flux, with the consequence that more than half of the resulting fluxes of the system became identifiable and indeed matched the original flux profile. Additional knowledge—or assumptions—about the fluxes can potentially constrain the feasible space of solutions further and may recover the original flux set. For example, knowing (or assuming) that a certain flux follows a specific functional form can potentially lead to a determination of this flux and decrease the degrees of freedom by one (cf. Iwata et al., 2013).

More generically, it is not always clear what optimality criteria or constraints should be evoked to reduce the feasible set of solutions, where all fit the concentration data exactly. Nonetheless, the identification and characterization of feasible flux sets may lead to a better understanding of the system and possibly aid the design of additional experiments that could effectively fill the gap and recover the true fluxes. Ideally, such experiments should yield data where all (most, or many) variables cover as much of their relevant substrate ranges as possible.

On a complementary trajectory, incomplete or missing data render the direct employment of DFE for the task of parameter

estimation impossible. Nonetheless, a mixed strategy of DFE and optimization may alleviate the problem and lead at least to subsets of identified fluxes.

AUTHOR CONTRIBUTIONS

SD and EV conceived the study. SD performed all analyses. SD and EV interpreted the results and wrote the paper.

ACKNOWLEDGMENTS

This work was supported in part by the following grants: NSF (MCB-0958172 and MCB-0946595; PI: EV; MCB 1411672; PI: Diana Downs); NIH (1P30ES019776-01A1, Gary W. Miller, PI); and DOE-BESC (DE-AC05-00OR22725; PI: Paul Gilna). BESC, the BioEnergy Science Center, is a U.S. Department of Energy Bioenergy Research Center supported by the Office of Biological and Environmental Research in the DOE Office of Science. The funding agencies are not responsible for the content of this article.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: <http://journal.frontiersin.org/article/10.3389/fgene.2016.00006>

REFERENCES

- Albert, A. (1972). *Regression and the Moore-Penrose Pseudoinverse*. New York, NY; London: Academic Press.
- Chou, I.-C., and Voit, E. O. (2009). Recent developments in parameter estimation and structure identification of biochemical and genomic systems. *Math. Biosci.* 219, 57–83. doi: 10.1016/j.mbs.2009.03.002
- Chou, I.-C., and Voit, E. O. (2012). Estimation of dynamic flux profiles from metabolic time series data. *BMC Syst. Biol.* 6:84. doi: 10.1186/1752-0509-6-84
- Curien, G., Bastien, O., Robert-Genthon, M., Cornish-Bowden, A., Cárdenas, M. L., Dumas, R., et al. (2009). Understanding the regulation of aspartate metabolism using a model based on measured kinetic parameters. *Mol. Syst. Biol.* 5:271. doi: 10.1038/msb.2009.29
- Dantzig, G. B. (1984). Reminiscences about the origins of linear-programming. *Mem. Am. Math. Soc.* 48, 1–11. doi: 10.1090/memo/0298
- Dolatshahi, S., Fonseca, L. L., and Voit, E. O. (2016a). New insights into the complex regulation of the glycolytic pathway in *Lactococcus lactis*. II. Inference of the precisely timed control system regulating glycolysis. *Mol. Biosyst.* 12, 37–47. doi: 10.1039/C5MB00726G
- Dolatshahi, S., Fonseca, L. L., and Voit, E. O. (2016b). New insights into the complex regulation of the glycolytic pathway in *Lactococcus lactis*. I. Construction and diagnosis of a comprehensive dynamic model. *Mol. Biosyst.* 12, 23–36. doi: 10.1039/C5MB00331H
- Dolatshahi, S., Vidakovic, B., and Voit, E. O. (2014). A constrained wavelet smoother for pathway identification tasks in systems biology. *Comput. Chem. Eng.* 71, 728–733. doi: 10.1016/j.compchemeng.2014.07.019
- Draper, N. R., and Smith, H. (1981). *Applied Regression Analysis, 2nd Edn*. New York, NY: Wiley. Series in Probability and Mathematical Statistics.
- Eilers, P. H. C. (2003). A perfect smoother. *Anal. Chem.* 75, 3631–3636. doi: 10.1021/ac034173t
- Goel, G., Chou, I.-C., and Voit, E. O. (2008). System estimation from metabolic time-series data. *Bioinformatics* 24, 2505–2511. doi: 10.1093/bioinformatics/btn470
- Gutenkunst, R. N., Casey, F. P., Waterfall, J. J., Myers, C. R., and Sethna, J. P. (2007a). Extracting falsifiable predictions from sloppy models. *Ann. N. Y. Acad. Sci.* 1115, 203–211. doi: 10.1196/annals.1407.003
- Gutenkunst, R. N., Waterfall, J. J., Casey, F. P., Brown, K. S., Myers, C. R., and Sethna, J. P. (2007b). Universally sloppy parameter sensitivities in systems biology models. *PLoS Comput. Biol.* 3:e189. doi: 10.1371/journal.pcbi.0030189
- Holzhütter, H. G. (2004). The principle of flux minimization and its application to estimate stationary fluxes in metabolic networks. *Eur. J. Biochem.* 271, 2905–2922. doi: 10.1111/j.1432-1033.2004.04213.x
- Iwata, M., Shiraishi, F., and Voit, E. O. (2013). Course but efficient identification of metabolic pathway systems. *Int. J. Syst. Biol.* 4, 57–72. Available online at: <http://www.bioinfopublication.org/viewhtml.php?artid=BIA0001914>
- Jia, G., Stephanopoulos, G., and Gunawan, R. (2011). Parameter estimation of kinetic models from metabolic profiles. *Bioinformatics* 27, 1964–1970. doi: 10.1093/bioinformatics/btr293
- Jia, G., Stephanopoulos, G., and Gunawan, R. (2012). Incremental parameter estimation of kinetic metabolic network models. *BMC Syst. Biol.* 6:142. doi: 10.1186/1752-0509-6-142
- Lewis, N. E., Hixson, K. K., Conrad, T. M., Lerman, J. A., Charusanti, P., Polpitiya, A. D., et al. (2010). Omic data from evolved *E. coli* are consistent with computed optimal growth from genome-scale models. *Mol Syst Biol.* 6:390. doi: 10.1038/msb.2010.47
- Liu, Y., and Gunawan, R. (2014). Parameter estimation of dynamic biological network models using integrated fluxes. *BMC Syst. Biol.* 8:127. doi: 10.1186/s12918-014-0127-x
- Marino, S., and Voit, E. O. (2006). An automated procedure for the extraction of metabolic network information from time series data. *J. Bioinform. Comput. Biol.* 4, 665–691. doi: 10.1142/S0219720006002259
- Penrose, R. (1955). A generalized inverse for matrices. *Math. Proc. Cambridge Philos. Soc.* 51, 406–413.
- Raue, A., Kreutz, C., Theis, F. J., and Timmer, J. (2013). Joining forces of bayesian and frequentist methodology: a study for inference in the presence of non-identifiability. *J. Philos. Trans. R. Soc. A* 371:20110544. doi: 10.1098/rsta.2011.0544

- Savageau, M. A. (1992). "Critique of the enzymologist's test tube," in *Fundamentals of Medical Cell Biology*, ed E. E. Bittar (Greenwich, CT: JAI Press Inc.), 45–108.
- Savageau, M. A. (1995). "Enzyme kinetics *in vitro* and *in vivo*: Michaelis-Menten revisited," in *Principles of Medical Biology*, ed E. E. Bittar (Greenwich, CT: JAI Press Inc.), 93–146.
- Savageau, M. A., and Voit, E. O. (1982). Power-law approach to modeling biological systems: I. Theory. *J. Ferment. Technol.* 60, 221–228.
- Siegenthaler, C., and Gunawan, R. (2014). Assessment of network inference methods. *PLoS ONE* 9:e90481. doi: 10.1371/journal.pone.0090481
- Sorribas, A., March, J., and Voit, E. O. (2000). Estimating age-related trends in cross-sectional studies using S-distributions. *Stat. Med.* 19, 697–713. doi: 10.1002/(SICI)1097-0258(20000315)19:5<697::AID-SIM378>3.0.CO;2-Y
- Tafintseva, V., Tøndel, K., Ponosov, A., and Martens, H. (2014). Global structure of sloppiness in a nonlinear model. *J. Chemomet.* 28, 645–655. doi: 10.1002/cem.2651
- Tönsing, C., Timmer, J., and Kreutz, C. (2014). Cause and cure of sloppiness in ordinary differential equation models. *Phys. Rev. E* 90:023303. doi: 10.1103/PhysRevE.90.023303
- Ud-Dean, S. M. M., and Gunawan, R. (2014). Ensemble inference and inferability of gene regulatory networks. *PLoS ONE* 9:e103812. doi: 10.1371/journal.pone.0103812
- Varah, J. M. (1982). A spline least squares method for numerical parameter estimation in differential equations. *SIAM J. Sci. Stat. Comput.* 3, 28–46. doi: 10.1137/0903003
- Vilela, M., Borges, C. C., Vinga, S., Vasconcelos, A. T., Santos, H., Voit, E. O., et al. (2007). Automated smoother for the numerical decoupling of dynamics models. *BMC Bioinformatics* 8:305. doi: 10.1186/1471-2105-8-305
- Vilela, M., Chou, I. C., Vinga, S., Vasconcelos, A. T., Voit, E. O., and Almeida, J. S. (2008). Parameter optimization in S-system models. *BMC Syst. Biol.* 2:35. doi: 10.1186/1752-0509-2-35
- Vilela, M., Vinga, S., Maia, M. A., Voit, E. O., and Almeida, J. S. (2009). Identification of neutral biochemical network models from time series data. *BMC Syst. Biol.* 3:47. doi: 10.1186/1752-0509-3-47
- Villaverde, A. F., and Banga, J. R. (2013). Reverse engineering and identification in systems biology: strategies, perspectives and challenges. *Interface* 11:20130505. doi: 10.1098/rsif.2013.0505
- Voit, E. O. (2012). *A First Course in Systems Biology*. New York, NY: Garland Science.
- Voit, E. O. (2000). *Computational Analysis of Biochemical Systems: A Practical Guide for Biochemists and Molecular Biologists*. Cambridge, UK: Cambridge University Press.
- Voit, E. O. (2011). "What if the fit is unfit? Criteria for biological systems estimation beyond residual errors," in *Applied Statistics for Biological Networks*, eds M. Dehmer, F. Emmert-Streib, and A. Salvador (New York, NY: J. Wiley and Sons), 183–200.
- Voit, E. O. (2013a). Biochemical systems theory: a review. *ISRN Biomath.* 2013, 53. doi: 10.1155/2013/897658
- Voit, E. O. (2013b). Characterizability of metabolic pathway systems from time series data. *Math. Biosci.* 246, 315–325. doi: 10.1016/j.mbs.2013.01.008
- Voit, E. O., and Almeida, J. (2004). Decoupling dynamical systems for pathway identification from metabolic profiles. *Bioinformatics* 20, 1670–1681. doi: 10.1093/bioinformatics/bth140
- Voit, E. O., Alvarez-Vasquez, F., and Hannun, Y. A. (2010). Computational analysis of sphingolipid pathway systems. *Adv. Exp. Med. Biol.* 688, 264–275. doi: 10.1007/978-1-4419-6741-1_19
- Voit, E. O., Goel, G., Chou, I. C., and Fonseca, L. L. (2009). Estimation of metabolic pathway systems from different data sources. *IET Systems Biol.* 3, 513–522. doi: 10.1049/iet-syb.2008.0180
- Voit, E. O., Martens, H. A., and Omholt, S. W. (2015). 150 Years of the mass action law. *PLoS Comp. Biol.* 11:e1004012. doi: 10.1371/journal.pcbi.1004012
- Voit, E. O., and Savageau, M. A. (1982a). Power-law approach to modeling biological systems; III. Methods of analysis. *J. Ferment. Technol.* 60, 223–241.
- Voit, E. O., and Savageau, M. A. (1982b). Power-law approach to modeling biological systems; II. Application to ethanol production. *J. Ferment. Technol.* 60, 229–232.
- Whittaker, E. T. (1923). On a new method of graduation. *Proc. Edinburgh Math. Soc.* 41, 63–75.

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2016 Dolatshahi and Voit. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



A Comparison of Deterministic and Stochastic Modeling Approaches for Biochemical Reaction Systems: On Fixed Points, Means, and Modes

Sayuri K. Hahl* and Andreas Kremling

Specialty Division for Systems Biotechnology, Faculty of Mechanical Engineering, Technische Universität München, Garching, Germany

OPEN ACCESS

Edited by:

Julio Vera González,
University Hospital Erlangen, Germany

Reviewed by:

Pengyue Zhang,
Indiana University Bloomington, USA
Thomas Millat,
University of Nottingham, UK

*Correspondence:

Sayuri K. Hahl
sayuri.hahl@tum.de

Specialty section:

This article was submitted to
Systems Biology,
a section of the journal
Frontiers in Genetics

Received: 15 January 2016

Accepted: 17 August 2016

Published: 31 August 2016

Citation:

Hahl SK and Kremling A (2016) A
Comparison of Deterministic and
Stochastic Modeling Approaches for
Biochemical Reaction Systems: On
Fixed Points, Means, and Modes.
Front. Genet. 7:157.
doi: 10.3389/fgene.2016.00157

In the mathematical modeling of biochemical reactions, a convenient standard approach is to use ordinary differential equations (ODEs) that follow the law of mass action. However, this deterministic ansatz is based on simplifications; in particular, it neglects noise, which is inherent to biological processes. In contrast, the stochasticity of reactions is captured in detail by the discrete chemical master equation (CME). Therefore, the CME is frequently applied to mesoscopic systems, where copy numbers of involved components are small and random fluctuations are thus significant. Here, we compare those two common modeling approaches, aiming at identifying parallels and discrepancies between deterministic variables and possible stochastic counterparts like the mean or modes of the state space probability distribution. To that end, a mathematically flexible reaction scheme of autoregulatory gene expression is translated into the corresponding ODE and CME formulations. We show that in the thermodynamic limit, deterministic stable fixed points usually correspond well to the modes in the stationary probability distribution. However, this connection might be disrupted in small systems. The discrepancies are characterized and systematically traced back to the magnitude of the stoichiometric coefficients and to the presence of nonlinear reactions. These factors are found to synergistically promote large and highly asymmetric fluctuations. As a consequence, bistable but unimodal, and monostable but bimodal systems can emerge. This clearly challenges the role of ODE modeling in the description of cellular signaling and regulation, where some of the involved components usually occur in low copy numbers. Nevertheless, systems whose bimodality originates from deterministic bistability are found to sustain a more robust separation of the two states compared to bimodal, but monostable systems. In regulatory circuits that require precise coordination, ODE modeling is thus still expected to provide relevant indications on the underlying dynamics.

Keywords: ordinary differential equations, chemical master equation, bistability, bimodality, gene expression, protein bursts

1. INTRODUCTION

In the last decades, the potential of mathematical modeling for the analysis of biological systems has widely been recognized. However, the reliability and explanatory power of such models depend greatly on the chosen modeling approaches, which may largely differ in several aspects like their level of detail or the approximations they are based on. This fact has led to debates not only with critics from other scientific fields, but also within the community of systems biologists (Gunawardena, 2014). Nowadays, modeling is still lacking any kind of gold standard, since it is highly specific toward the underlying systems biological problem. In fact, each and every model can only provide a rough depiction of nature, and a major challenge consists in applying or even developing modeling techniques which answer the questions to be addressed as best as possible with reasonable effort.

The host of existing modeling approaches can be grouped according to various criteria. One important classification distinguishes between deterministic and stochastic models. In deterministic modeling, stochasticity within the system is neglected. One of the most frequently used deterministic approaches consists in ordinary differential equations (ODEs), which are based on the phenomenological law of mass action. They provide a dynamic and quantitative description of spatially homogenous systems. Since ODEs are intensely used in other scientific fields as well, numerous analysis techniques and simulation methods have been developed thus far. In theoretical biology, ODEs have been applied to a wide range of problems, for example to the description of metabolism (Kremling et al., 2007), signaling (Shinar et al., 2007) or gene regulation within cells (Tyson and Othmer, 1978), to the investigation of systemic effects in complex multicellular organisms (Gallenberger et al., 2012), and to the analysis of population dynamics (Edelstein-Keshet, 1988).

However, biological systems are always subject to stochastic effects, which occur on all levels—from molecular to macroscopic. These can be captured by stochastic models. Concerning biochemical networks, the chemical master equation (CME) is very frequently applied (van Kampen, 2007, Chapter 5). Unfortunately, its analytical solution is usually intractable, especially if a large number of reactants is involved. The Gillespie algorithm provides exact simulations of trajectories of the master equation (Gillespie, 1977), but the computational cost is high for multi-component reaction systems. Therefore, several approximate variants of the CME as well as of the Gillespie algorithm have been developed (Gillespie, 2001; Chatterjee et al., 2005; Anderson, 2008).

Randomness plays a major role in signaling and regulation, where the copy number of the involved components is small and noise in gene expression is significant. Therefore, they are major application fields for stochastic models in systems biology (Tsimring, 2014). Compared to their deterministic counterparts, stochastic models are in general more difficult to analyze. Therefore, the need for incorporating stochasticity should be carefully elucidated, depending on the biological application.

In the following, we aim at comparing the explanatory power of the very detailed discrete CMEs and the corresponding ODEs. Starting from reviewing the analogies in their formulation in Section 2, we will then collect the parallels and discrepancies between the modeling results in Section 3. In this context, common concepts like bistability and bimodality will be contrasted. Unlike a couple of other studies on this topic, we will also regard mesoscopic systems which are not close to the thermodynamic limit. We will discuss these aspects in the context of a simple gene regulatory system, using it as a platform for identifying general factors which influence the comparability between these kinds of deterministic and stochastic models.

2. THEORETICAL BACKGROUND

2.1. Foundations of CMEs and ODEs

In this section, we will review the formulation of CMEs and ODEs for chemical reaction systems, and highlight the connection between deterministic reaction rates and stochastic reaction propensities. More detailed descriptions can be found, e.g., in Gillespie (2007) and van Kampen (2007).

2.1.1. Chemical Reactions as a Markov Process and the Chemical Master Equation

Let us consider a system containing molecules of M different chemical species (components) that can in total undergo R different irreversible, elementary reactions. These reactions can be of zeroth order (e.g., entry of molecules into an open system), of first order (e.g., degradation of compounds or unimolecular conversion) or of higher order (e.g., dimerization). In the latter case, random encounters of two or more molecules are necessary for the reaction to occur. The j -th reaction can be written as

$$\sum_{i=1}^M \beta_{ij} \cdot X_i \longrightarrow \sum_{i=1}^M \gamma_{ij} \cdot X_i, \quad (1)$$

with X_i denoting the components in the system and $\beta_{ij}, \gamma_{ij} \in \mathbb{N}_0^+$ being the stoichiometric coefficients of the educts and products. Assuming that the system is spatially homogeneous, its state can be characterized by the copy numbers of the components it contains. It can therefore be formulated in terms of a vector $\mathbf{n}(t) = (n_1(t), \dots, n_M(t))^T$, where n_i denotes the copy number of the i -th component, and t is the time variable.

In the CME framework, the system state is modeled as a continuous-time stochastic process, for which the Markov property holds. This means that the probability distribution of future system states only depends on the present state, but not on past states (memorylessness). Here, we regard the discrete state space defined above. Let $p_{\mathbf{n}}(t)$ be the probability of being in state \mathbf{n} at time t and $\pi(\mathbf{n}, \mathbf{m})$ be the probability per infinitesimal time unit (propensity) of a transition from \mathbf{m} to \mathbf{n} . The CME is a reformulation of the Chapman-Kolmogorov equation and can be written as:

$$\dot{p}_{\mathbf{n}}(t) = \sum_{\mathbf{m}} (\pi(\mathbf{n}, \mathbf{m}) p_{\mathbf{m}}(t) - \pi(\mathbf{m}, \mathbf{n}) p_{\mathbf{n}}(t)), \quad (2)$$

where $\dot{p}_{\mathbf{n}}(t)$ denotes the time derivative of the probability and the summation runs over the whole state space. The CME thus states that the temporal evolution of $p_{\mathbf{n}}$ is determined by the balance between transitions leading to state \mathbf{n} and transitions away from \mathbf{n} . Since Equation (2) applies to all states \mathbf{n} , it defines a system of differential equations describing the dynamics of the probability mass function p .

Next, $\pi(\cdot, \cdot)$ needs to be defined in the context of the reaction system (Equation 1). Within infinitesimal intervals, transitions occur solely due to single reactions. The stoichiometric matrix \mathbf{A} with entries $a_{ij} := \gamma_{ij} - \beta_{ij}$ and columns $\mathbf{a}_j = (a_{1j}, \dots, a_{Mj})^\top$ captures all possible transitions between states, so that the CME can be re-formulated as

$$\dot{p}_{\mathbf{n}}(t) = \sum_{j=1}^R (w_j(\mathbf{n} - \mathbf{a}_j) \cdot p_{\mathbf{n}-\mathbf{a}_j}(t) - w_j(\mathbf{n}) \cdot p_{\mathbf{n}}(t)). \quad (3)$$

Here, $w_j(\mathbf{n})$ is the propensity of the j -th reaction, which is the probability per infinitesimal time unit for the j -th reaction to occur, when the system is in state \mathbf{n} . The propensities can more specifically be formulated as

$$w_j(\mathbf{n}) = \kappa_j \cdot \prod_{i=1}^M \binom{n_i}{\beta_{ij}}. \quad (4)$$

Here, κ_j denotes the stochastic reaction constant, which is determined by physical properties of the reaction (e.g., activation energy, complexity) and by environmental conditions like temperature. The latter product reflects the combinatorial probability of random encounters of the educts: it accounts for reactive collisions of the components, where β_{ij} out of n_i molecules of the i -th component are involved.

2.1.2. Formulation of a System of Ordinary Differential Equations

We consider ODEs based on the law of mass action, which has originally been formulated by Guldberg and Waage. In this deterministic approach, concentrations instead of molecule numbers are usually regarded (Gillespie, 1976), and the state space is treated as continuous (Gillespie, 1977). This is only justified if the molecule numbers of the species and the system size V are sufficiently large. Let $c_i := \frac{n_i}{V}$ be the concentration of the i -th reaction component. For constant V , the concentration change of the i -th component through the reactions in Equation (1) is given by

$$\dot{c}_i = \sum_{j=1}^R \left((\gamma_{ij} - \beta_{ij}) \cdot k_j \cdot \prod_{l=1}^M c_l^{\beta_{lj}} \right) = \sum_{j=1}^R \left(a_{ij} \cdot k_j \cdot \prod_{l=1}^M c_l^{\beta_{lj}} \right). \quad (5)$$

k_j is the deterministic rate constant. The law of mass action thus states that the speed of a reaction depends on this constant and on powers of the concentrations of the educts. In case of elementary reactions, the exponents are determined by the stoichiometry of the reaction.

2.1.3. Relation between Stochastic and Deterministic Reaction Constants

While the stochastic reaction constant reflects the likelihood of a reaction to occur, the deterministic counterpart is mostly interpreted as a kinetic term. However, the following mathematical relation holds:

$$\kappa_j = k_j \cdot V \cdot \prod_{i=1}^M \frac{\beta_{ij}!}{V^{\beta_{ij}}}. \quad (6)$$

This equation is a generalized form of the relation derived in Gillespie (1977). The stochastic rate constant thus depends on the system size V , and this dependence is determined by the stoichiometry: While zero-order reactions are more likely to happen in large systems, the chance of molecular collisions required for higher-order reactions is reduced when the density of molecules decreases due to an expansion of V . Inserting the relation into Equation (4) yields

$$w_j(\mathbf{n}) = \frac{k_j}{V^{\sum_{i=1}^M \beta_{ij}-1}} \prod_{i=1}^M \frac{n_i!}{(n_i - \beta_{ij})!}. \quad (7)$$

For small and well-characterized chemical reaction systems, the CME and ODE formulations are straightforward. However, many biological reactions in cellular systems are complex and a description in terms of elementary reactions like in Equation (1) might thus be difficult. For example, the conversion of a protein with the help of an enzyme is actually a multi-step process. The description of gene expression including the actions of the transcription and translation machinery and under the influence of certain activators or repressors would be infeasible at this level of detail. By exploiting time scale separation, it is a general practice to lump several fast reactions into rate functions $k_j(\mathbf{n})$, which replace the constants k_j and which depend on the current system state (pseudo-steady-state assumption). They can for example be chosen to describe Hill-type kinetics.

2.2. The Mean of the CME and Its Relation to the ODE System

The deterministic formulation is sometimes regarded as a description of average values, which are assumed to represent the system quite well when the molecule numbers of the components and the system size are large. However, a basic calculation of the mean of the CME shows that this analogy only holds true in special cases (cf. for example van Kampen, 2007, Chapter 5):

Let N_i be the stochastic variable of the copy number of the i -th component and let $\mathbf{N} = (N_1, \dots, N_M)^\top$. Let furthermore $\mathbb{E}[\cdot]$ denote the expected value. Then, the ODE of $\mathbb{E}[N_i]$ satisfies:

$$\dot{\mathbb{E}}[N_i] = \sum_{\mathbf{n} \in \mathbb{Z}^M} n_i \dot{p}_{\mathbf{n}} = \sum_{j=1}^R (a_{ij} \cdot \mathbb{E}[w_j(\mathbf{N})]). \quad (8)$$

For the sake of simplicity, we have omitted the time variable t . The derivation can be found in the Supplementary Material 1. Inserting the explicit formulation of the propensities (Equation 7) leads to

$$\dot{\mathbb{E}}[N_i] = \sum_{j=1}^R \left(a_{ij} \cdot \mathbb{E} \left[\frac{k_j}{V^{\sum_{l=1}^M \beta_{lj}-1}} \prod_{l=1}^M \frac{N_l!}{(N_l - \beta_{lj})!} \right] \right). \quad (9)$$

If all $\beta_{ij} \leq 1$, or if the system is close to the thermodynamic limit (i.e., the theoretical limit $V \rightarrow \infty$, $n_l \rightarrow \infty$ s.t. c_l is constant), the approximation $\frac{n_l!}{(n_l - \beta_{ij})! V^{\beta_{ij}}} \approx \frac{n_l^{\beta_{ij}}}{V^{\beta_{ij}}} = c_l^{\beta_{ij}}$ holds.

The expectation of the random variable $C_i := \frac{N_i}{V}$ describing the concentration then reads:

$$\mathbb{E}[C_i] = \sum_{j=1}^R \left(a_{ij} \cdot \mathbb{E} \left[k_j \prod_{l=1}^M C_l^{\beta_{lj}} \right] \right). \quad (10)$$

In general, $\mathbb{E}[f(Y)] \neq f(\mathbb{E}[Y])$ for any nonlinear function f , where Y is an arbitrary random variable. A comparison between the ODE in Equation (10) and the deterministic formulation in Equation (5) thus shows that the deterministic variable c_i is only an exact description of $\mathbb{E}[C_i]$, if the term $k_j \prod_{l=1}^M C_l^{\beta_{lj}}$ is linear. This holds true for zero and first order reactions with constant k_j , which is quite a severe restriction in the context of biochemical processes.

2.3. Bistability vs. Bimodality

In addition to the calculations in the preceding section, one further, quite obvious limitation in identifying deterministic variables with the mean of stochastic variables becomes apparent when multimodal probability distributions are regarded. They have more than one local maximum, each of them representing a characteristic composition of state variables that is “favored” by the system. Multimodality therefore reflects system state heterogeneity. This heterogeneity might be temporal (frequent switching of individual systems between different states) or population-based (split of a population into subgroups with different, but stable characteristics). If deterministic models were a mere description of the mean, they would obscure this multimodal structure and would therefore be rather uninformative. Indeed, a property of ODE models exists which describes some sort of heterogeneity: This property is called multistability, meaning that multiple stable fixed points can be assumed by the system. The initial condition determines which of the states the system will finally tend to. Although the effect of stochasticity, which might allow for random transitions between the stable states, is neglected, multistability has long been regarded as the deterministic equivalent of multimodality.

Recently, several theoretical as well as experimental studies have challenged this association. Bistable systems with a unimodal distribution have been observed as well as bimodal systems whose deterministic description predicted monostability (Artyomov et al., 2007; Qian et al., 2009; Bishop and Qian, 2010; Ochab-Marcinek and Tabaka, 2010; To and Maheshri, 2010; Shu et al., 2011; McSweeney and Popovic, 2014). We can thus conclude that deterministic variables are neither fully equivalent to the stochastic mean nor to stochastic modes. This raises the question under which conditions deterministic models can provide reliable information on system dynamics and which qualitative and quantitative conclusions can be drawn from the results.

In Gillespie (2007, 2009), Kurtz (1972, 1980), and van Kampen (2007), connections between deterministic and stochastic variables have been derived which are valid in the

thermodynamic limit under certain constraints on the reaction system. These constraints are usually fulfilled for elementary reactions, but might be violated when multiple reactions are lumped together. Furthermore, in gene expression and regulation, where the molecule copy numbers of some of the involved components are low, the thermodynamic limit is not an appropriate approximation.

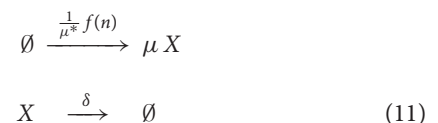
In order to characterize possible differences between ODE and CME models in a mesoscopic regime, we regard a flexible biochemical regulatory system that can be bimodal, depending on the parameters. Usually, bimodality arises due to positive feedback loops—a topological structure which can be found in autostimulatory gene expression systems, both natural and synthetic. They offer a fruitful platform for studying the effect of stochasticity in a biological context: Protein production was found to occur in bursts of random size, which enables us to study the influence of stochasticity and stoichiometry by varying the burst characteristics, namely the average amplitude and frequency. Moreover, by theoretically varying the feedback structure, the effect of nonlinear reaction rates can be studied. Using this system, we will determine in which aspects and to what extent the deterministic description is consistent with the CME. More general statements will then be derived from our observations.

3. RESULTS

3.1. Modeling of a Gene Regulatory System with Feedback

Our basic description of gene regulation is mainly adapted from Friedman et al. (2006) and Aquino et al. (2012). Instead of modeling the dynamics on the promoter, mRNA, and protein level in detail, we will use a simplification proposed in Aquino et al. (2012), which regards only protein formation and degradation. It is based on a time scale separation with subsequent averaging of promoter states and of mRNA concentrations. A discussion on the validity of this approximation is also given in Aquino et al. (2012). We will not put special emphasis on the accuracy of the model from a biological point of view, but rather on the mathematical characteristics of the model equations. We thus prefer the reduced model due to its analytical solvability.

The reaction scheme we consider is given by:



Here, X denotes the protein, which is generated in a burst with random size $\mu \in \mathbb{N}^+$. The burst size follows a geometric distribution with mean μ^* . The rate of protein production is given by $f: \mathbb{R} \rightarrow \mathbb{R}$, which is a smooth monotonically increasing function evaluated at the integer protein molecule number n , illustrating autostimulation. It is scaled by μ^* in order to obtain comparable results when the parameter is modified in our analyses (i.e., a change of the burst size is balanced by a

reciprocal change of the burst frequency so that the mean protein level remains constant). The protein degradation rate is linear with parameter δ . Throughout this study, the cell volume V is assumed to be fixed, so that dilution effects are neglected.

This scheme is suitable for studying the effect of linear as well as of nonlinear functions f , by which different types of autoregulation can be represented. For example, a non-cooperative stimulatory effect of the protein on its own expression can be described by a linearly increasing function or by a Michaelis-Menten-type saturation function. Cooperative feedback, where several protein molecules exert a synergistic autoregulatory effect, can be described through a sigmoid Hill-type function. Furthermore, by choosing large values for μ^* , significant jumps in the protein trajectories can be generated.

3.2. Deterministic Description in Terms of Ordinary Differential Equations

Since stochasticity is neglected in deterministic descriptions, the number of proteins produced in each burst is assumed to be equal to the average burst size μ^* . Let $c := \frac{n}{V}$ be the protein concentration, which is treated as a continuous variable. The ODE is then given by:

$$\dot{c} = \mu^* \frac{1}{\mu^*} \tilde{f}(c) - \delta c = \frac{\tilde{f}(c)}{V} - \delta c, \quad (12)$$

where $\tilde{f}(c) := f(c \cdot V) \forall c \in \mathbb{R}$. The steady state condition reads

$$\frac{\tilde{f}(c)}{\delta} = c V. \quad (13)$$

The number of fixed points thus depends on the structure of $\frac{\tilde{f}(c)}{\delta}$ and can be determined graphically as shown in the top row of **Figure 1**: The red line corresponds to the left hand-side of Equation (13) and has therefore the shape of \tilde{f} , and the identity line marked in green depicts the right hand-side. The steady states are located at the intersection points. Provided that the basal rate of protein production is nonzero, systems without feedback (panel A) or with non-cooperative positive feedback (panel B) can only have one stable fixed point. Those two cases are modeled by constant and by monotonically increasing, concave \tilde{f} , respectively. In case of cooperative feedback, which is characterized by a sigmoid structure of \tilde{f} , the system can be mono- or bistable (panels C and D).

3.3. Mathematical Formulation Using the Chemical Master Equation

The master equation of the reaction system is obtained by inserting the rates and stoichiometry given in Equation (11) into Equation (3). It can be written as:

$$\dot{p}_n = \sum_{\mu=1}^n \left(g_{\mu^*}(\mu) \frac{1}{\mu^*} f(n-\mu) p_{n-\mu} \right) - \frac{1}{\mu^*} f(n) p_n + \delta(n+1) p_{n+1} - \delta n p_n. \quad (14)$$

$g_{\mu^*}(\mu) := \frac{1}{\mu^*} \cdot \left(1 - \frac{1}{\mu^*}\right)^{\mu-1}$, $\mu \in \mathbb{N}^+$, is the geometric probability mass function.

According to the calculation in Aquino et al. (2012) using Z-transform (a discrete version of the Laplace transform), the probability mass function in steady state ($\dot{p}_n = 0 \forall n$) can be formulated recursively as

$$p_1^{\text{ss}} = \frac{f(0)}{\delta \mu^*} p_0^{\text{ss}},$$

$$(n+1) p_{n+1}^{\text{ss}} = \frac{f(n)}{\delta \mu^*} p_n^{\text{ss}} + \frac{\mu^* - 1}{\mu^*} n p_n^{\text{ss}}. \quad (15)$$

3.4. Calculation of Central Moments and Modes

The ODEs for the expectation and the variance σ^2 of the master equation read:

$$\frac{d\mathbb{E}[N]}{dt} = \mathbb{E}[f(N)] - \delta \mathbb{E}[N] \quad (16)$$

$$\frac{d\sigma^2(N)}{dt} = 2 \text{Cov}(N, f(N)) - (\mathbb{E}[f(N)] - \delta \mathbb{E}[N]) + 2 \mu^* \mathbb{E}[f(N)] - 2 \delta \sigma^2(N), \quad (17)$$

where N is the discrete random variable of the number of protein molecules. The detailed calculation is shown in the Supplementary Material 2. In steady state, the central moments are therefore given by:

$$\mathbb{E}[N] = \frac{\mathbb{E}[f(N)]}{\delta} \quad (18)$$

$$\sigma^2(N) = \mu^* \mathbb{E}[N] + \frac{1}{\delta} \text{Cov}(N, f(N)). \quad (19)$$

Hence, the variance depends on the mean burst size μ^* . For example, if f was constant (no feedback), $\sigma^2(N) = \mu^* \mathbb{E}[N]$ holds.

Let us now focus on the extrema of the probability distribution. In general, local maxima (modes) and minima obey the following conditions:

$$p_{n-1}^{\text{ss}} \leq p_n^{\text{ss}}, \quad p_n^{\text{ss}} \geq p_{n+1}^{\text{ss}} \rightarrow \text{maximum at } n \quad (20)$$

$$p_{n-1}^{\text{ss}} \geq p_n^{\text{ss}}, \quad p_n^{\text{ss}} \leq p_{n+1}^{\text{ss}} \rightarrow \text{minimum at } n, \quad (21)$$

if $n > 0$. Furthermore, one extremum necessarily occurs at $n = 0$.

Using Equation (15), one obtains the specific condition:

$$p_{n+1}^{\text{ss}} - p_n^{\text{ss}} \stackrel{\leq}{\geq} 0 \Leftrightarrow \frac{f(n)}{\delta} \stackrel{\leq}{\geq} n + \mu^*. \quad (22)$$

Thus, the extrema satisfy the condition

$$n = \left\lceil \frac{f(n)}{\delta} - \mu^* \right\rceil, \quad (23)$$

where $\lceil \cdot \rceil$ denotes the ceiling function.

3.5. Comparison of the Deterministic and Stochastic Descriptions

A comparison of the differential Equations (12) and (16) shows that the average (scaled by the volume) deviates from the deterministic variable if $\mathbb{E}[f(N)] \neq f(\mathbb{E}[N])$, which is usually the case when f is nonlinear. Inserting the Taylor series of $f(N)$ around $\mathbb{E}[N]$, the ODE of the mean reads:

$$\frac{d\mathbb{E}[N]}{dt} = f(\mathbb{E}[N]) + \sum_{r=2}^{\infty} \left(\frac{1}{r!} z_r f^{(r)}(\mathbb{E}[N]) \right) - \delta \mathbb{E}[N] \quad (24)$$

with $z_r := \mathbb{E}[(N - \mathbb{E}[N])^r]$ denoting the r -th central moment of N and $f^{(r)}$ being the r -th derivative of f . The mean of the CME is thus well described by the deterministic variable c only if f is almost linear or if higher central moments of N like the variance, skewness, kurtosis, etc. are small. As already mentioned, Equation (19) shows that bursting leads to a significant increase in the variance. Taken together, nonlinearity in the reaction can cause a deviation of $c \cdot V$ from $\mathbb{E}[N]$, which is expected to be enhanced through bursting.

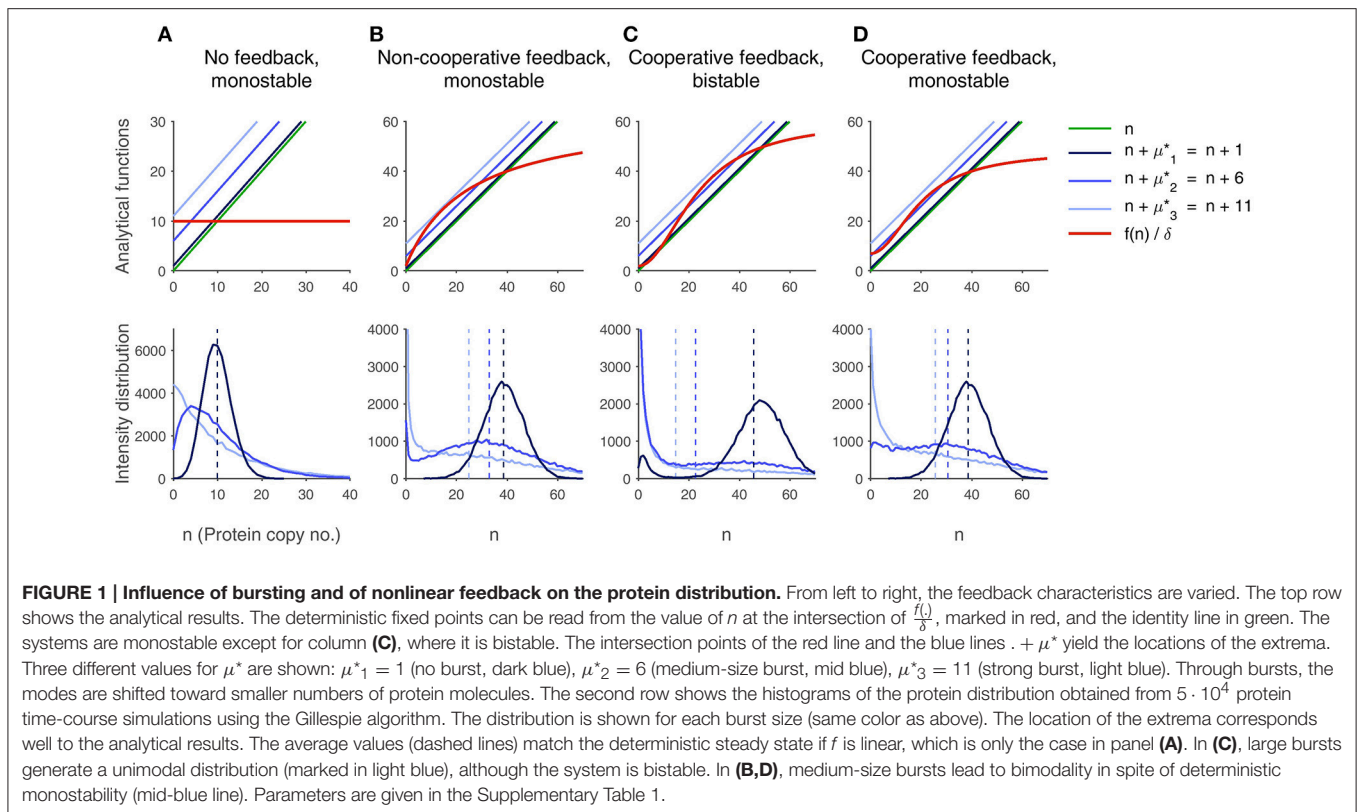
Concerning the modes, a comparison of the conditions given in Equation (13) and in Equation (23) reveals a strong analogy if μ^* is small, so that stable fixed points can be associated with the maxima in the equilibrium probability mass function, and the unstable fixed points correspond to the minima. However, large bursts can disrupt this connection, as will be shown in the following section. The structure of f plays a minor role in this context.

3.5.1. Large Protein Bursts Can Disrupt the Connection between Bistability and Bimodality

In **Figure 1**, our previous calculations are visualized. Protein time-courses have been simulated using the Gillespie algorithm. For each plot, $5 \cdot 10^4$ simulations were run and the histograms at a final time point t_f were plotted. In order to make sure that the steady state was approximately reached, several runs with random initial molecule numbers have been performed and compared to one another, and the simulated means and modes have been compared to the analytical values.

The first row of plots illustrates the analytical results, summarizing the findings from deterministic fixed point analysis as well as from the calculation of the stochastic extrema: According to Equation (13), the deterministic fixed points can be read from the intersection points of the graphs of $\frac{f(n)}{\delta}$ and n . The approximate location of the modes is given by the intersection of $\frac{f(n)}{\delta}$ and $n + \mu^*$, see Equation (23). From left to right, the structure of f was changed in order to check different feedback mechanisms. Furthermore, the mean burst size was varied in each case: Missing bursts ($\mu_1^* = 1$), medium-size bursts ($\mu_2^* = 6$) and large bursts ($\mu_3^* = 11$) were considered. The plots below show the corresponding histograms of the simulations for the three different burst sizes. Moreover, the empirical mean of the distribution is highlighted.

First, let $f \equiv b$ be constant. The simulations in panel (A) show that an increase in μ^* does not change the location of the empirical mean. However, the maximum of the distribution



is biased toward smaller values, while the variance is enlarged. These observations are in perfect agreement to our calculations: The fixed point is located at $\frac{b}{\delta}$ and matches the empirical mean due to the linearity of f . The variance is given by $\sigma^2(N) = \mu^* \frac{b}{\delta}$, it therefore depends on the burst size. The mode fulfills the condition $n = \left\lceil \frac{b}{\delta} - \mu^* \right\rceil$ and is thus shifted to the left when μ^* is increased.

If f is a non-cooperative saturation curve (panel B), the deterministic steady state deviates from the empirical average of the distribution, and the bias is enlarged under bursting conditions, as stated before. Furthermore, the fixed point only matches with the maximum of the histogram when bursts are very small. Very interestingly, μ^* can even be large enough to generate a bimodal distribution which peaks at $n = 0$ and at a positive value. If the burst size is further increased, the distribution can eventually turn unimodal again, the only maximum being located at zero. This is also predicted by our analytical considerations, where the shift of the identity line by μ^* leads to the emergence of another intersection point with $\frac{f(n)}{\delta}$, corresponding to the formation of a minimum in the distribution, and a further shifting makes the intersection points vanish, so that the only maximum is found at $n = 0$. As a consequence, bursting can cause bimodality although the deterministic description predicts monostability.

Panel (C) addresses sigmoid functions f , which are often the result of protein oligomerizations leading to cooperative behavior. First, we have chosen the parameters such that the deterministic system is bistable. Again, an increase in μ^* shifts the modes to the left so that the deviation of the deterministic steady states increases. Under large bursts, as in the non-cooperative case, a unimodal distribution peaking at $n = 0$ can be observed. This is an example for a bistable system, which is unimodal.

By varying the parameters in the system with cooperative feedback, the results shown in panel (D) are obtained. The system is monostable, but it can get bimodal under bursting conditions. In contrast to the situation shown in (B), both maxima are located at positive molecule numbers.

All in all, large and rare bursts lead to an asymmetry in the protein production and degradation events, generating a skewed probability density with a large variance, that cannot be approximated by a normal distribution (cf. the association of deterministic and stochastic models via the Langevin equation in Gillespie, 2007). This disrupts the connection between deterministic fixed points and stochastic modes.

3.5.2. Good Agreement between the Stochastic and Deterministic Descriptions in the Thermodynamic Limit

In spite of the preceding results, which reveals the possibility of huge deviations between the outcome of deterministic and stochastic models, the following calculation shows that in the theoretical thermodynamic limit $V \rightarrow \infty$, $n \rightarrow \infty$, s.t. $\frac{n}{V}$ is

constant, a strong correlation between modes and fixed points exists.

Let us consider a system whose size is increased s -fold compared to system (Equation 11) (i.e., its volume is given by $s \cdot V$). In order to maintain the concentrations, the rate of translation, which is formally a zeroth-order reaction, needs to be increased accordingly (it is thus given by $s \cdot \frac{1}{\mu^*} \tilde{f}$). In this case, the deterministic ODE remains unchanged, since Equation (12) is simply replaced by the identical formulation

$$\dot{c} = \frac{s \tilde{f}(c)}{s V} - \delta c. \quad (25)$$

The condition for the stochastic modes reads

$$n = c s V = \left\lceil \frac{s \cdot \tilde{f}(c)}{\delta} - \mu^* \right\rceil \quad (26)$$

$$\Rightarrow c V = \frac{\left\lceil \frac{s \cdot \tilde{f}(c)}{\delta} - \mu^* \right\rceil}{s} \xrightarrow{s \rightarrow \infty} \frac{\tilde{f}(c)}{\delta} \quad (27)$$

and thus matches the deterministic fixed point in the thermodynamic limit. The simulations shown in **Figure 2**, where protein distributions of two systems with differing sizes are compared, confirm this result. To put it in a slightly different way, the modes are in good agreement with the deterministic steady states if μ^* is small relative to the value of n at the extremum. However, note that from merely locating the deterministic fixed points in a bistable system, one cannot infer the average steady-state of the system, since the probability of a cell to be in one or the other state is unknown.

3.6. Feedback and Burst Characteristics Influence the Precision of the Distribution and the Robustness of Bimodality

Next, we will give a qualitative estimate on the local precision (i.e., the inverse of the variance) of the probability distribution at the modes. The recursive formula (15) can be written as

$$p_{n+1}^{ss} - p_n^{ss} = \frac{\frac{f(n)}{\delta} - (n + \mu^*)}{\mu^*(n + 1)} p_n^{ss}. \quad (28)$$

Therefore, the local change of the probability mass function relative to its height is large if

- $\left| \frac{f(n)}{\delta} - (n + \mu^*) \right|$ is large, while
- μ^* is small.

Under this condition, the local distribution forms sharp peaks around a maximum located at n . As a consequence, feedback structures and burst characteristics have a significant impact on the separation of the modes (without loss of generality, we do not include the effect of the degradation rate δ into our considerations). In the following, three scenarios will be portrayed which illustrate this result. They are visualized in **Figure 3**.

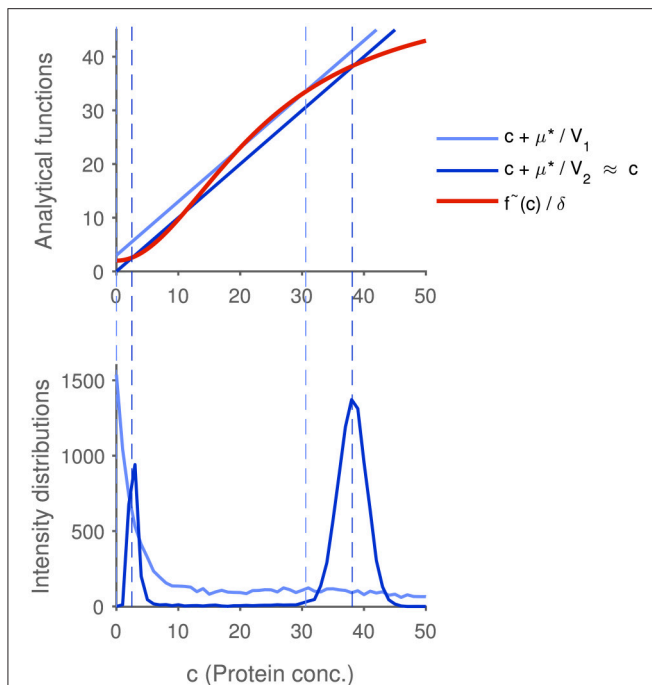


FIGURE 2 | Influence of the system size on the correspondence between deterministic and stochastic modeling results.

Two systems with differing sizes are compared: The volume V_1 of system 1 (graphs in light blue) is chosen 50-fold smaller than the volume V_2 of system 2 (graphs in dark blue), while the protein concentrations at the deterministic fixed points are identical. The intersections of the blue lines and the red line in the upper plot mark the analytical locations of the extrema in the protein probability mass function. The extrema of the larger system nearly coincide with the deterministic fixed points, since the expression $\frac{\mu^*}{V_2}$ is almost negligible. The distributions in the bottom plot (obtained using the Gillespie algorithm) confirm these results: the larger system shows a clear bimodal distribution whose modes match the stable deterministic fixed points, while the modes of the small system are shifted, and the distribution is much broader. The dashed lines show that the analytical determination of the modes fits well to the simulations. Parameters are given in the Supplementary Table 2.

First, the burst size μ^* is varied, while $\frac{f(n)}{\delta} - (n + \mu^*)$ and the location of the modes are kept constant. This is most easily achieved by choosing two different burst sizes μ_1^* and μ_2^* with $\mu_1^* < \mu_2^*$ and by defining $f_2 := f_1 + \delta(\mu_2^* - \mu_1^*)$ (i.e., the system with the larger average burst size has an enhanced basal protein production, while the shapes (derivatives) of f_1 and f_2 are identical). Then, $\frac{f_1(n)}{\delta} - (n + \mu_1^*) = \frac{f_2(n)}{\delta} - (n + \mu_2^*)$ holds true. The simulated histogram in **Figure 3A** shows that the system with the larger burst size does indeed have a broader distribution.

Next, the function f is modified while μ^* is held fixed. In this context, cooperative and non-cooperative regulation can be compared (cf. **Figure 3B**): Let $f_3(n) := b_3 + v_3 \frac{n^h}{n^h + K_3}$ with $h > 1$ describe Hill type kinetics (sigmoid function, cooperativity) and let $f_4(n) := b_4 + v_4 \frac{n}{n + K_4}$ be a Michaelis-Menten-type function (no cooperativity). Furthermore, to ensure comparability, let $b_3 = b_4$ (identical basal expression level), and let $\frac{f_3}{\delta}$ and $\frac{f_4}{\delta}$ have two identical intersection points with the line $\cdot + \mu^*$. This guarantees that the modes of the probability distributions occur at the same

protein molecule numbers. Then, it can be shown that for all n , $\left| \frac{f_4(n)}{\delta} - (n + \mu^*) \right| < \left| \frac{f_3(n)}{\delta} - (n + \mu^*) \right|$ holds (the proof is given in the Supplementary Material 3), so that the protein distribution of the cooperative system has sharper peaks.

Finally, both μ^* and $\frac{f(n)}{\delta} - (n + \mu^*)$ are varied. Let $f_5(n) := b_5 + v_5 \frac{n}{n + K_5}$, $f_6(n) := b_6 + v_6 \frac{n}{n + K_6}$, and let $b_5 = b_6$. Now, let $\mu_5^* > \mu_6^*$ and let $\frac{f_5(n)}{\delta} = n + \mu_5^*$ and $\frac{f_6(n)}{\delta} = n + \mu_6^*$ have an identical set of solutions. We are thus looking at two non-cooperative systems where the basal rate of protein production and the locations of the modes coincide, while the burst sizes and the curvatures of f_5 and f_6 differ. In this case, $\left| \frac{f_5(n)}{\delta} - (n + \mu_5^*) \right| > \left| \frac{f_6(n)}{\delta} - (n + \mu_6^*) \right|$ (cf. Supplementary Material 4), which counteracts the effect of the differing burst sizes. Explicit calculations are therefore required to determine which effect prevails. Interestingly, **Figure 3C** shows that the bimodality in the protein distribution of the system with larger bursts is even more precise.

Having addressed the probability mass function in steady state, single protein time-courses are now regarded. In a bimodal system, the robustness of the two stable steady states is crucial for its functionality: The protein level might fluctuate permanently between these states (small Mean first-passage times (MFPTs) of transitions between the inactive and active states, cf. van Kampen, 2007, Chapter 12), or it might tend to stay in one of the states with rare switching events (large MFPTs). The trajectories in **Figure 3** show that a sharp bimodal distribution qualitatively correlates well with the robustness of the states. In **Figure 3A**, the fluctuations in the system with the lower burst level are much smaller, leading to more distinct switches between the modes. The protein level of the system with cooperative feedback in **Figure 3B** has small noise and stays in the active state, whereas the protein time-course in the non-cooperative circuit does not exhibit a clear separation of the modes. The time-courses in **Figure 3C** show that even systems with non-cooperative regulation are able to sustain two separate states, given that the nonlinearity of the feedback and the burst size are not too small, which severely contradicts the results of standard deterministic modeling.

4. DISCUSSION

In this study, we have compared an ODE model based on the law of mass action with the corresponding CME formulation, implicitly stating that the master equation provides the much more realistic description of the biochemical reaction system. All deviations of the deterministic from the stochastic model have thus been interpreted as an indication of inadequacy of the ODE formalism. Indeed, as Gillespie states, “the stochastic approach is always valid whenever the deterministic approach is valid, and is sometimes valid when the deterministic approach is not” (Gillespie, 1976). One should still note that the CME, too, is based on several simplifying assumptions. Among these are the random, homogenous distribution of positions AND velocities of reactants, which is only a valid approximation when elastic molecular collisions predominate over reactive ones (Nicolis et al., 1974; Gillespie, 1992). Hence, we need to point out

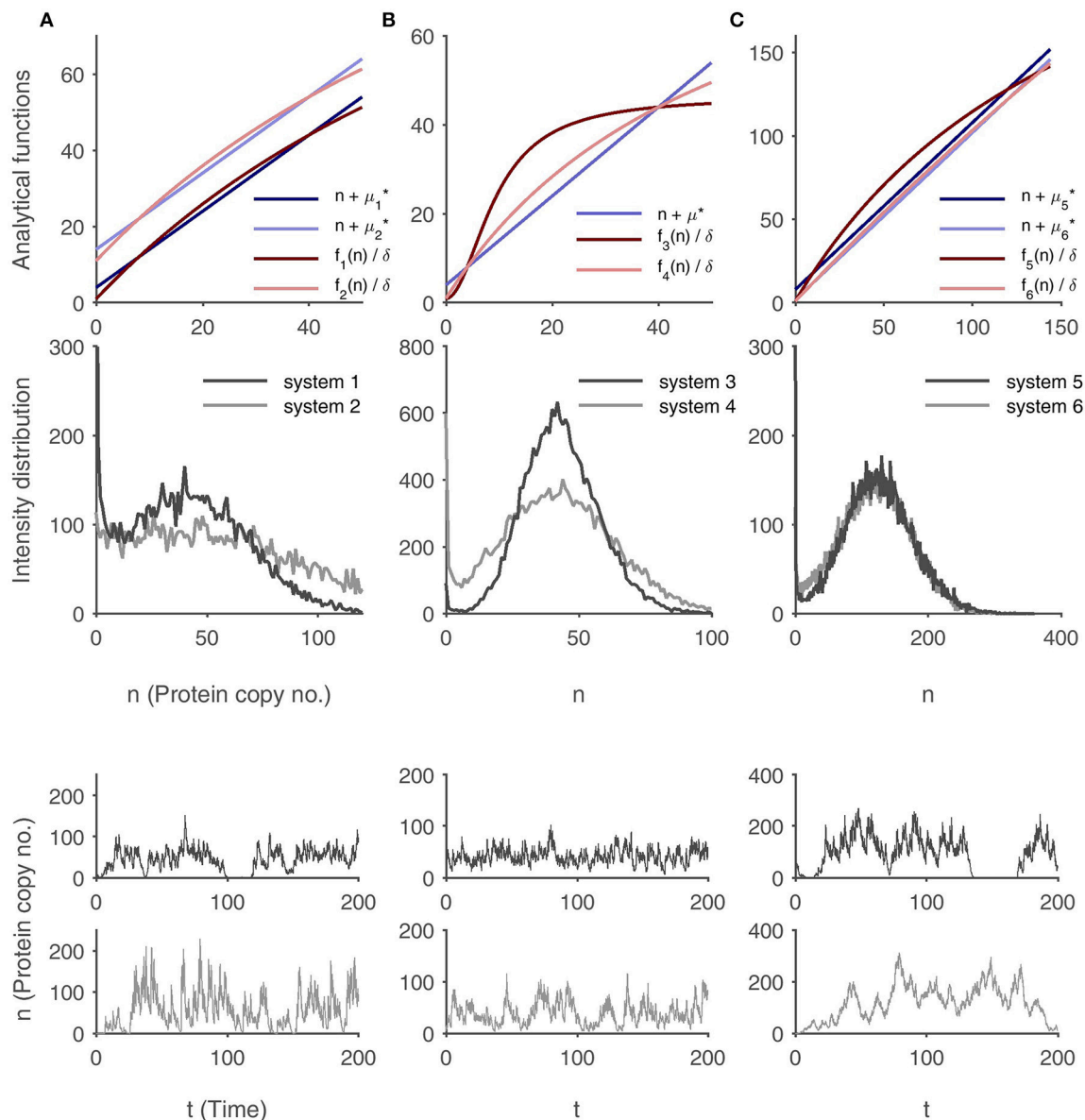


FIGURE 3 | Robustness of bimodality in different regulatory systems with feedback. In each column, the robustness of the modes in two regulatory systems with varying burst sizes and varying functions f are compared based on the protein distributions and exemplary protein time-courses. **(A)** Comparison of two non-cooperative feedback regulations with differing burst sizes $\mu_1^* < \mu_2^*$ and accordingly shifted functions f_1 and f_2 with identical shape. **(B)** Comparison of non-cooperative and cooperative regulation with identical burst size μ^* . **(C)** Comparison of two non-cooperative regulations with identical basal protein expression and under differing burst sizes $\mu_6^* < \mu_5^*$. In all cases, the system marked in dark colors (system 1, 3, and 5, respectively) exhibits a sharper distribution and a better separation of the modes is visible in the protein time-course simulations. Further explanations are given in the main text. Parameter values are listed in the Supplementary Table 3.

that although the CME approach often leads to experimentally verifiable results, this cannot be taken for granted. On the other hand, we can state that if significant mathematical deviations of the even more simplistic ODE approach from the CME model are observed, the deterministic description is almost surely unrealistic. Our study has led to the conclusion that although ODE modeling is quite a convenient and popular approach in many application fields, the use of deterministic models should

be treated cautiously in the context of mesoscopic biochemical reaction systems.

The connection between deterministic and stochastic modeling has frequently been studied before. Several papers have reported on multi-component reaction systems that are monostable and bimodal, where bimodality is caused by the presence of components with very slow dynamics. These components can act as multi-level switches on fast downstream

components (Qian et al., 2009; Ochab-Marcinek and Tabaka, 2010; Shu et al., 2011). Here, we have focused on nonlinear one-component reaction systems. A related study was previously conducted by Bishop and Qian (2010), where a phosphorylation-dephosphorylation cycle has been analyzed. They have shown that although the one-dimensional deterministic ODE model exhibits monostability, the weak nonlinearity in the reactions has the potential to cause stochastic bimodality, if the system size is sufficiently small. In their case, one of the stationary modes was invariably located at the zero state, whereas the other one was close to the deterministic steady state.

Here, we have systematically analyzed the effects of nonlinearity, but also of large stoichiometric coefficients in a flexible autoregulated gene expression system. In this context, we have proposed a graphical method which visualizes the impact of these system properties on the location of the modes and on their deviation from the deterministic fixed points. With the help of the graphics, it could be shown that monostable but bimodal systems can be constructed with both modes occurring at positive values, but only if the feedback is cooperative. We have seen that large stoichiometric coefficients can promote highly asymmetric, irregular fluctuation patterns in the copy numbers of the components. In our example, protein bursts allow for sudden and large increases in the number of protein molecules, whereas single degradation events reduce the number merely by one. Such instant jumps in molecule numbers have been explicitly excluded in the publications by Gillespie, where deterministic and stochastic variables were found to correspond well in sufficiently large systems (Gillespie, 2007, 2009). We have shown that when all reactions are linear, the mean and the deterministic variable coincide, but skewed fluctuations through large bursts lead to a shift of the mode away from the mean. In the presence of nonlinear reaction propensities, the deterministic variable usually differs from the mean, and large bursts can even qualitatively change the modality of the distribution. One could argue that through a more detailed description of the bursting mechanism, large stoichiometric coefficients can to some extent be avoided. Nevertheless, there are components within a cell which usually occur at single-digit amounts (e.g., genes, mRNA), so that every reaction involving them is inevitably accompanied by a “large jump” relative to

the molecule number. As a next step, the interplay of jumps, nonlinearities and reaction time-scales in a multi-component reaction system needs to be evaluated. Our preliminary results (not shown) indicate that those three factors together can further reduce the comparability of ODE and CME models.

This provokes the question of what kind of conclusions can still be drawn from deterministic modeling in small-scale reaction systems. In some biological contexts, stochasticity plays an important functional role: noise in certain signaling and gene regulation systems can lead to random transitions between different stable state and thus serve to create population heterogeneity, which makes cells more robust toward fluctuating environmental conditions. In this case, deterministic trajectories are certainly not realistic. But often, uniform cellular behavior can be observed. A coordinated hysteretic switch from one state to another, for example, is only possible if the modes are robustly separated. We have shown that although monostable systems can be bimodal with moderate switching frequency, a more robust bimodality is generated in a regime which is indeed deterministically bistable. In such cases, deterministic modeling might still provide valuable information on the dynamics of the system. For a more reliable description of biochemical processes in mesoscopic systems, however, we think that the use of stochastic modeling is virtually inevitable.

AUTHOR CONTRIBUTIONS

AK supervised the project; AK and SH designed research; SH carried out analysis, simulations, and interpretation; All authors wrote and approved the manuscript.

ACKNOWLEDGMENTS

SH is funded by the German Ministry of Education and Research (BMBF) through the e:bio initiative (grant number: 031A299C).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: <http://journal.frontiersin.org/article/10.3389/fgene.2016.00157>

REFERENCES

- Anderson, D. F. (2008). Incorporating postleap checks in tau-leaping. *J. Chem. Phys.* 128:054103. doi: 10.1063/1.2819665
- Aquino, T., Abranches, E., and Nunes, A. (2012). Stochastic single-gene autoregulation. *Phys. Rev. E Stat. Nonlin. Soft Matter Phys.* 85(6 Pt 1):061913. doi: 10.1103/PhysRevE.85.061913
- Artyomov, M. N., Das, J., Kardar, M., and Chakraborty, A. K. (2007). Purely stochastic binary decisions in cell signaling models without underlying deterministic bistabilities. *Proc. Natl. Acad. Sci. U.S.A.* 104, 18958–18963. doi: 10.1073/pnas.0706110104
- Bishop, L. M., and Qian, H. (2010). Stochastic bistability and bifurcation in a mesoscopic signaling system with autocatalytic kinase. *Biophys. J.* 98, 1–11. doi: 10.1016/j.bpj.2009.09.055
- Chatterjee, A., Vlachos, D. G., and Katsoulakis, M. A. (2005). Binomial distribution based tau-leap accelerated stochastic simulation. *J. Chem. Phys.* 122:024112. doi: 10.1063/1.1833357
- Edelstein-Keshet, L. (1988). *Mathematical Models in Biology*. New York, NY: McGraw-Hill.
- Friedman, N., Cai, L., and Xie, X. S. (2006). Linking stochastic dynamics to population distribution: an analytical framework of gene expression. *Phys. Rev. Lett.* 97:168302. doi: 10.1103/PhysRevLett.97.168302
- Gallenberger, M., zu Castell, W., Hense, B. A., and Kuttler, C. (2012). Dynamics of glucose and insulin concentration connected to the β -cell cycle: model development and analysis. *Theor. Biol. Med. Model.* 9:46. doi: 10.1186/1742-4682-9-46
- Gillespie, D. T. (1976). A general method for numerically simulating the stochastic time evolution of coupled chemical reactions.

- J. Comput. Phys. 22, 403–434. doi: 10.1016/0021-9991(76)90041-3
- Gillespie, D. T. (1977). Exact stochastic simulation of coupled chemical reactions. *J. Phys. Chem.* 81, 2340–2361. doi: 10.1021/j100540a008
- Gillespie, D. T. (1992). A rigorous derivation of the chemical master equation. *Phys. A Stat. Mech. Appl.* 188, 404–425. doi: 10.1016/0378-4371(92)90283-V
- Gillespie, D. T. (2001). Approximate accelerated stochastic simulation of chemically reacting systems. *J. Chem. Phys.* 115:1716. doi: 10.1063/1.1378322
- Gillespie, D. T. (2007). Stochastic simulation of chemical kinetics. *Annu. Rev. Phys. Chem.* 58, 35–55. doi: 10.1146/annurev.physchem.58.032806.104637
- Gillespie, D. T. (2009). Deterministic limit of stochastic chemical kinetics. *J. Phys. Chem. B* 113, 1640–1644. doi: 10.1021/jp806431b
- Gunawardena, J. (2014). Models in biology: ‘accurate descriptions of our pathetic thinking’. *BMC Biol.* 12:29. doi: 10.1186/1741-7007-12-29
- Kremling, A., Bettenbrock, K., and Gilles, E. D. (2007). Analysis of global control of *Escherichia coli* carbohydrate uptake. *BMC Syst. Biol.* 1:42. doi: 10.1186/1752-0509-1-42
- Kurtz, T. G. (1972). The relationship between stochastic and deterministic models for chemical reactions. *J. Chem. Phys.* 57:2976.
- Kurtz, T. G. (1980). *Biological Growth and Spread, volume 38 of Lecture Notes in Biomathematics*. Berlin; Heidelberg: Springer.
- McSweeney, J. K., and Popovic, L. (2014). Stochastically-induced bistability in chemical reaction systems. *Ann. Appl. Probab.* 24, 1226–1268. doi: 10.1214/13-AAP946
- Nicolis, G., Allen, P., and Van Nypelseer, A. (1974). Some remarks on the theory of fluctuations around nonequilibrium states. *Prog. Theor. Phys.* 52, 1481–1497. doi: 10.1143/PTP.52.1481
- Ochab-Marcinek, A., and Tabaka, M. (2010). Bimodal gene expression in noncooperative regulatory systems. *Proc. Natl. Acad. Sci. U.S.A.* 107, 22096–22101. doi: 10.1073/pnas.1008965107
- Qian, H., Shi, P.-Z., and Xing, J. (2009). Stochastic bifurcation, slow fluctuations, and bistability as an origin of biochemical complexity. *Phys. Chem. Chem. Phys.* 11, 4861–4870. doi: 10.1039/b900335p
- Shinar, G., Milo, R., Martínez, M. R., and Alon, U. (2007). Input output robustness in simple bacterial signaling systems. *Proc. Natl. Acad. Sci. U.S.A.* 104, 19931–19935. doi: 10.1073/pnas.0706792104
- Shu, C.-C., Chatterjee, A., Dunny, G., Hu, W.-S., and Ramkrishna, D. (2011). Bistability versus bimodal distributions in gene regulatory processes from population balance. *PLoS Comput. Biol.* 7:e1002140. doi: 10.1371/journal.pcbi.1002140
- To, T.-L., and Maheshri, N. (2010). Noise can induce bimodality in positive transcriptional feedback loops without bistability. *Science* 327, 1142–1145. doi: 10.1126/science.1178962
- Tsimring, L. S. (2014). Noise in biology. *Rep. Prog. Phys.* 77:026601. doi: 10.1088/0034-4885/77/2/026601
- Tyson, J. J., and Othmer, H. G. (1978). “The dynamics of feedback control circuits in biochemical pathways,” in *Progress in Theoretical Biology*, Vol. 5, eds F. Snell and R. Rosen (New York, NY: Academic Press), 2–62.
- van Kampen, N. G. (2007). *Stochastic Processes in Physics and Chemistry, 3rd Edn.* Amsterdam: North Holland.

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2016 Hahl and Kremling. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



Multiplicity of Mathematical Modeling Strategies to Search for Molecular and Cellular Insights into Bacteria Lung Infection

Martina Cantone[†], Guido Santos[†], Pia Wentker, Xin Lai and Julio Vera^{*}

Laboratory of Systems Tumor Immunology, Department of Dermatology, Friedrich-Alexander University Erlangen-Nürnberg and Universitätsklinikum Erlangen, Erlangen, Germany

OPEN ACCESS

Edited by:

Matteo Barberis,
University of Amsterdam, Netherlands

Reviewed by:

Frederick R. Adler,
University of Utah, United States

Reiko J. Tanaka,
Imperial College London,
United Kingdom

*Correspondence:

Julio Vera
julio.vera-gonzalez@uk-erlangen.de

[†]These authors have contributed
equally to this work.

Specialty section:

This article was submitted to
Systems Biology,
a section of the journal
Frontiers in Physiology

Received: 03 March 2017

Accepted: 16 August 2017

Published: 30 August 2017

Citation:

Cantone M, Santos G, Wentker P,
Lai X and Vera J (2017) Multiplicity of
Mathematical Modeling Strategies to
Search for Molecular and Cellular
Insights into Bacteria Lung Infection.
Front. Physiol. 8:645.
doi: 10.3389/fphys.2017.00645

Even today two bacterial lung infections, namely pneumonia and tuberculosis, are among the 10 most frequent causes of death worldwide. These infections still lack effective treatments in many developing countries and in immunocompromised populations like infants, elderly people and transplanted patients. The interaction between bacteria and the host is a complex system of interlinked intercellular and the intracellular processes, enriched in regulatory structures like positive and negative feedback loops. Severe pathological condition can emerge when the immune system of the host fails to neutralize the infection. This failure can result in systemic spreading of pathogens or overwhelming immune response followed by a systemic inflammatory response. Mathematical modeling is a promising tool to dissect the complexity underlying pathogenesis of bacterial lung infection at the molecular, cellular and tissue levels, and also at the interfaces among levels. In this article, we introduce mathematical and computational modeling frameworks that can be used for investigating molecular and cellular mechanisms underlying bacterial lung infection. Then, we compile and discuss published results on the modeling of regulatory pathways and cell populations relevant for lung infection and inflammation. Finally, we discuss how to make use of this multiplicity of modeling approaches to open new avenues in the search of the molecular and cellular mechanisms underlying bacterial infection in the lung.

Keywords: systems biology, systems medicine, lung infection, mathematical modeling, Boolean network, ODE models, stochastic modeling, agent-based modeling

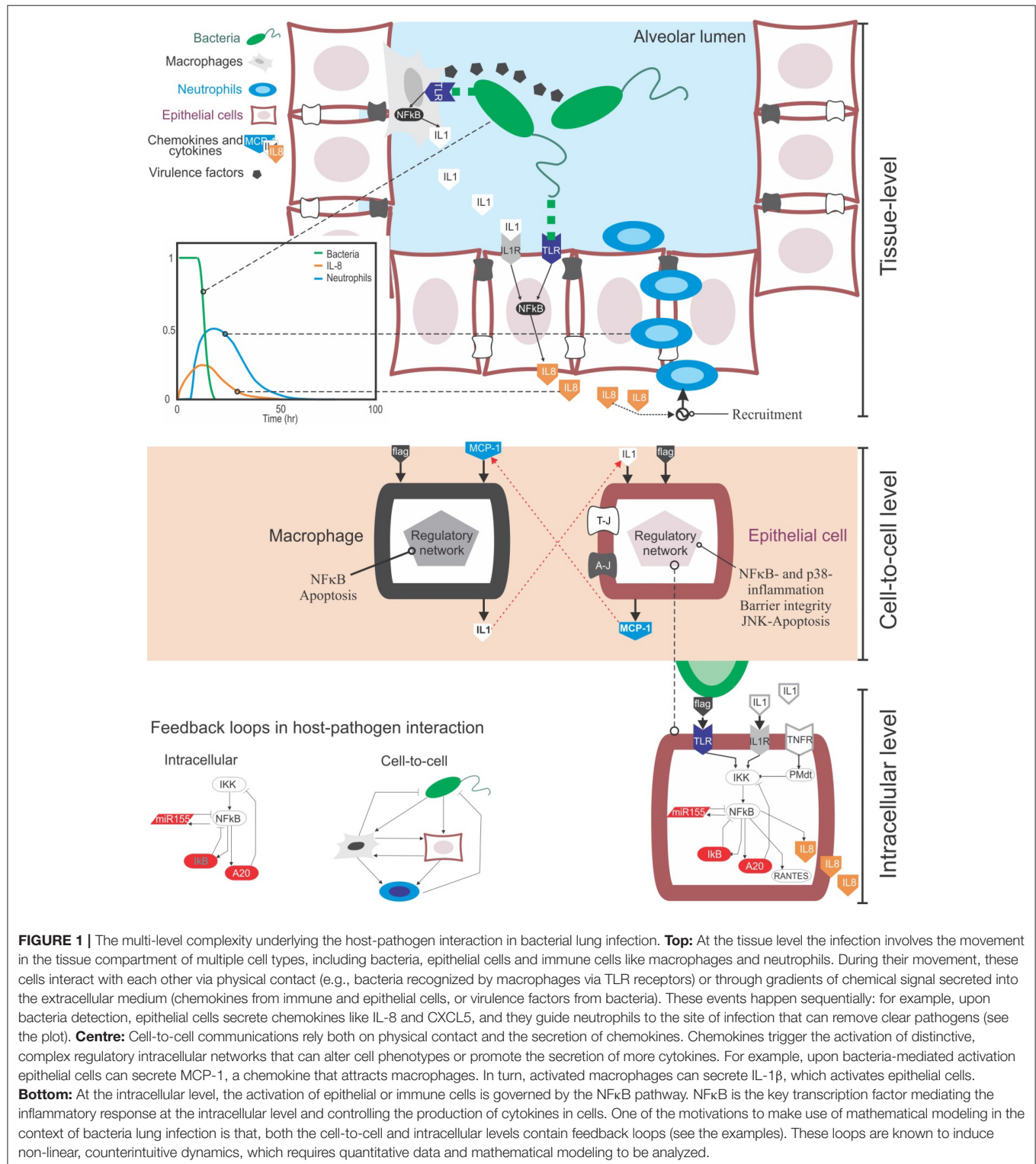
INTRODUCTION

In a time of moon shooting projects to cure cancer (Nature Editorial, 2016), the reader may wonder why it remains interesting to deploy a “systemic approach” to deepen our understanding of bacterial lung infections. First, even nowadays two of the 10 most frequent causes of death worldwide are bacterial infections targeting the lungs, namely pneumonia and tuberculosis (WHO, 2017b). A few generations ago, respiratory infections used to claim the life of a significant fraction of infants, a problem circumvented in western countries with the emergence of antibiotics, sulfonamides and high quality health care, but still a dramatic reality in many developing countries. Second, elderly individuals and immunocompromised individuals face the challenge of repeated respiratory infections (Stupka et al., 2009). A similar problem is faced by immunocompromised populations (Conces, 1998).

Finally, bacteria resistant to antibiotics create new risks and motivate the struggle to create new antibiotics (Silver, 2011; WHO, 2017a).

Bacteria and other microbes can invade the lung through the airways. When pathogens reach the lumen of lung alveoli

they can replicate and attack the tissue using virulence factors, their own chemical weaponry (**Figure 1**). Upon recognition of pathogens, the immune response is initiated to clear them from the infected sites, and this process involves the secretion of cytokines and recruitment of immune cells.



A balanced immune response can be achieved via interacting immune cells that are controlled by intracellular regulatory networks of interacting molecules, such as cytokines, receptors, kinases, transcription factors, or non-coding RNAs. Such a system contains regulatory motifs, especially positive and negative feedback loops, which increase the complexity of the response and can provoke non-linear behaviors such as bistability and oscillation (Ref). For patients with respiratory bacterial infections, severe pathological condition can emerge if their immune systems fail to quickly neutralize the infection and to avoid systemic spread of the pathogen. On the other hand, overwhelming host immune response to the pathogens is also dangerous and can impede the proper functioning of the lung and other organs. So, any new treatments using the combination of antibiotics and immunomodulatory drugs will be useful if they can help the patients to maintain a balanced immune response (Wentker et al., 2017), which is governed by the multi-level biological system (Eberhardt et al., 2016).

This level of complexity is equivalent to other natural and artificial systems, like those controlling large and modern aircrafts. For decades researchers in physics and engineering have been using mathematical modeling and simulations as an irreplaceable tool when trying to understand, predict or redesign these systems. Systems Medicine is the natural extension of this strategy to the biomedical domain. In our context, mathematical modeling can be used: (a) to inspect and integrate different but complementary types of quantitative experimental and clinical data, (b) to design experiments, (c) to elaborate, analyze and discuss hypotheses, (d) to perform model simulation-based predictions for the course of a disease, or (e) the feasibility of conventional, newly developed or personalized treatments (Vera and Wolkenhauer, 2008). For our purposes, Systems Medicine is a methodology that employs mathematical modeling to integrate and analyze quantitative biological data (Auffray et al., 2009; Wolkenhauer et al., 2013; Eberhardt et al., 2016; **Figure 2**). In the approach, biological knowledge is encoded into mathematical models whose simulations are used to dissect the cellular and molecular mechanisms behind diseases.

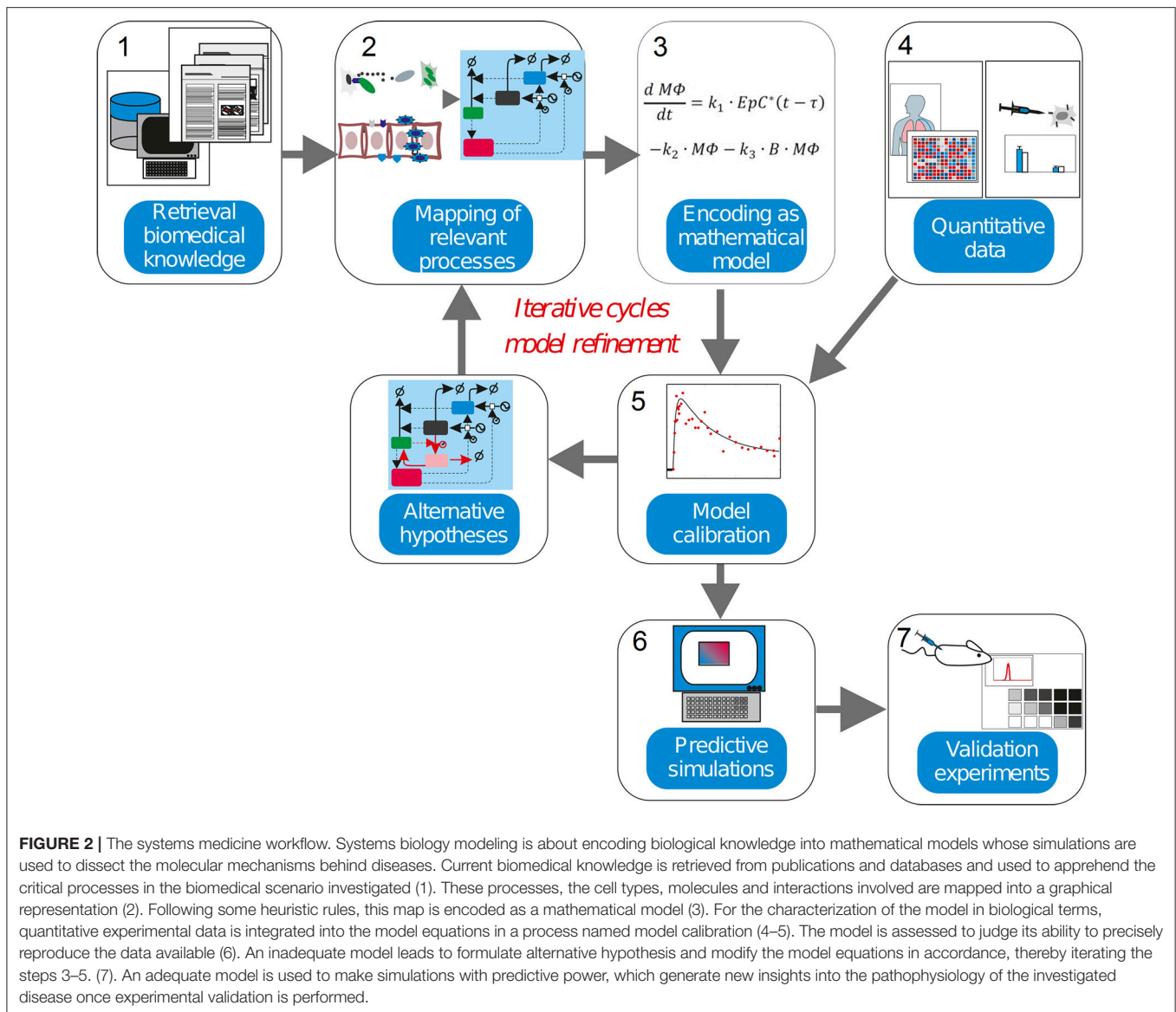
In a nutshell, the workflow is composed of several steps (**Figure 2**). The model derivation begins by **retrieval biomedical knowledge (1)**, biomedical information from publications and databases is used to identify the key compounds (cell types or molecules) and their interactions, and translated the information into a graphical depiction named regulatory map, **mapping of relevant processes (2)**. Based on the information gathered and some heuristic rules, this map is **encoded as a mathematical model (3)**, which consists of equations or other mathematical entities. In **model calibration (4-5)**, quantitative data obtained from experiments are used to characterize the mathematical model. This is often done through a computational process called “model calibration,” which assigns values to the parameters characterizing the model equations, such as the model becomes able to reproduce the existing quantitative data. Model calibration can often confirm or disprove the hypothesis encoded by the model equations. The inability of the mathematical model to reproduce the data leads to its reformulation, and eventually to the design of new experiments. In **predictive simulations**

(6), a calibrated model is used to generate new insights into the pathophysiology of the investigated disease via computer simulation. Finally, further **validation experiments (7)** are used to confirm or discard the predictions made via model simulation.

In the same manner as one cannot elucidate all the mysteries of modern biomedicine using a single experimental technique, say confocal microscopy, a single class of mathematical model among the plethora of those available in systems medicine is not useful for every purpose. Every problem or hypothesis to be explored requires a carefully selected and specific modeling approach. In this paper, we discuss and illustrate the distinctive features of different mathematical modeling frameworks with cases studies in the context of bacterial lung infection. Further, we compile and discuss relevant published results on the mathematical modeling of pathways and networks modulating the immune response, the host-pathogen interaction and the occurrence of coinfections, all of them topics relevant for bacterial lung infection. Finally, we discuss how to make use of this multiplicity of modeling approaches to open new avenues in the search of molecular and cellular insights in bacterial lung infection. This review is intended for modelers who want to enter the field of bacterial lung infection and need a review of published work, but also for infectiologists and immunologists interested on understanding how mathematical modeling can help them designing and interpreting their quantitative data and hypothesis. In the main text we focus on the basis of the modeling workflow, the modeling approaches and the published results, while further details in the methodologies discussed and the examples proposed are provided in Supplementary Material.

MATHEMATICAL MODELING OF BACTERIAL LUNG INFECTION

In the context of lung infection, the use of mathematical modeling is especially suited because one is interested on elucidating the function and regulation of cell-to-cell or biochemical networks governing the local or systemic activation of the epithelial and immune cells in the course of lung bacterial infection. These networks are large and tightly interconnected; further, they display complex patterns of temporal activation. Moreover, one can be interested on integrating quantitative clinical and biological data accounting for the dynamics of the infection across different time- and spatial scales. Some events triggering the early local lung infection happens within minutes to hours, while the systemic phase of the immune response and the recovery and tissue repair can last days to weeks. Something similar happens at the spatial organization, with microscopy-level events like the triggering intracellular networks or the networks of interacting immune cells at the infection site, and mesoscopic-level events accounting for the effects of infection in the make-up and functioning of structures the lung alveoli and the airways. This level of complexity in terms of structure and data can be managed using different types of mathematical modeling. In the following we discuss in detail several modeling approaches, as well as the context during bacterial lung infection in which they are valid.



BOOLEAN MODELS

Main Features of Boolean Models

Biochemical systems, if treated as networks of interacting entities, share many of the structural and regulatory features of electronic circuits. Boolean models, conceived for designing electronic circuits, were proposed 50 years ago as a tool to investigate the structure and dynamics of biochemical networks (Kauffman, 1969, 1993). For biochemical systems, Boolean networks are graphs in which nodes represent molecules and edges represent interactions between molecules. The interplay between molecules and biochemical reactions is represented using Boolean logic, i.e., discrete models in which every node or molecule can have only binary values: 0 or “OFF” (indicating the nonexistence or no-activation of the considered biochemical species), and 1 or “ON” (corresponding to its existence or activation). For example,

Figure 3B is a depiction of the activation of the IL-1 β receptor (IL-1 β R) upon binding of its ligand (IL-1 β). The process can be modeled using a Boolean logic function “AND.” The table in **Figure 3C** represents all the possible combinations for the values of IL-1 β and IL-1 β R and their effect in the values of the activated receptor IL-1 β R*. One can see that activation (IL-1 β R* “1”) is only possible if IL-1 β R and IL-1 β are present (both with value “1”).

In Boolean Networks (BNs), the set of functions used to represent interactions is reduced to the basic logic gates “AND-OR-NOT” (for definition of logic gate and any blue marked word, see Glossary Section). However, logic gates can be combined in multiple ways and therefore complex multi-molecule interactions can be represented (Shmulevich and Aitchison, 2009; Wang et al., 2012). In line with this, the intracellular regulatory networks underlying the activation of immune cells can be investigated

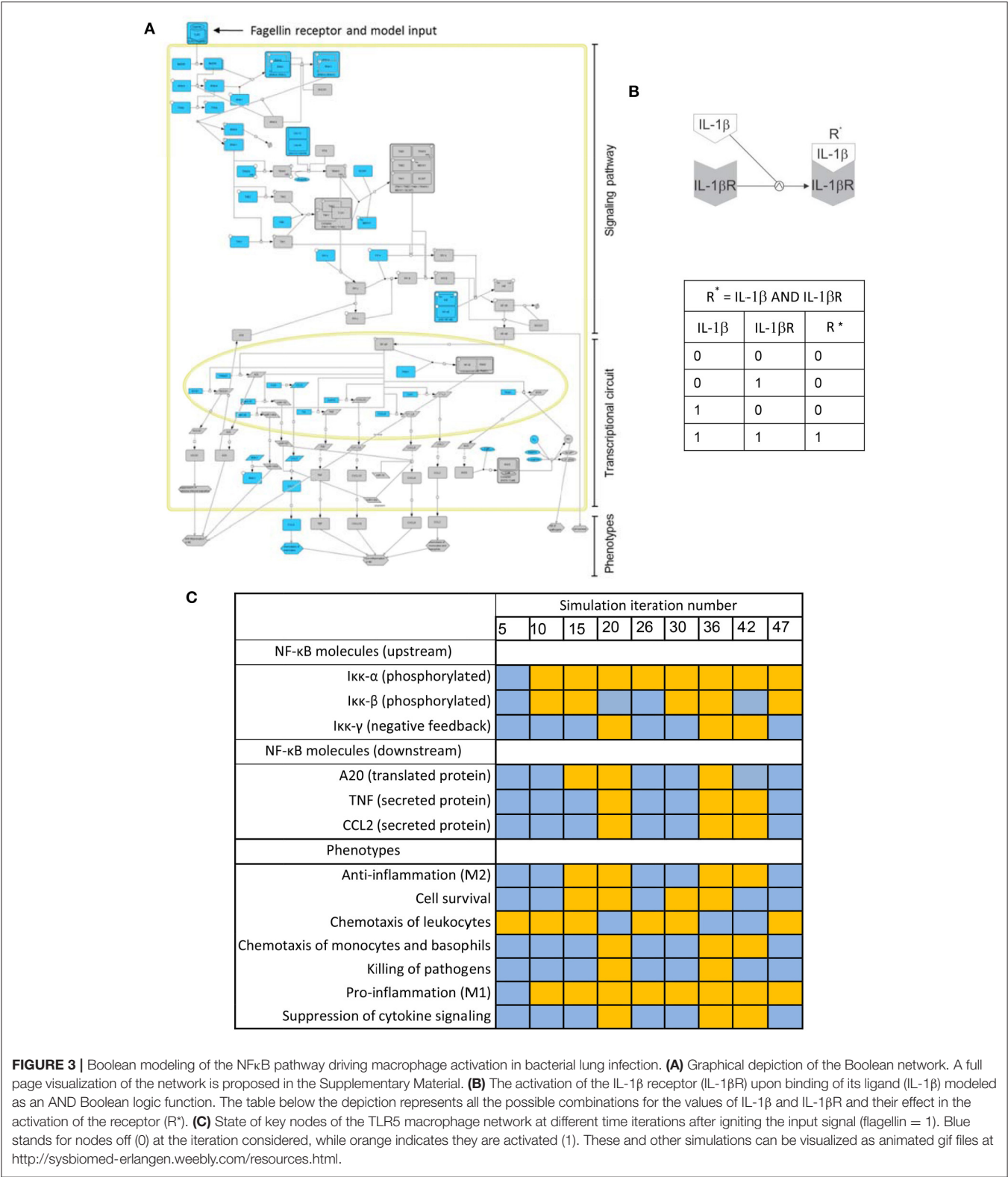


FIGURE 3 | Boolean modeling of the NFκB pathway driving macrophage activation in bacterial lung infection. **(A)** Graphical depiction of the Boolean network. A full page visualization of the network is proposed in the Supplementary Material. **(B)** The activation of the IL-1β receptor (IL-1βR) upon binding of its ligand (IL-1β) modeled as an AND Boolean logic function. The table below the depiction represents all the possible combinations for the values of IL-1β and IL-1βR and their effect in the activation of the receptor (R*). **(C)** State of key nodes of the TLR5 macrophage network at different time iterations after igniting the input signal (flagellin = 1). Blue stands for nodes off (0) at the iteration considered, while orange indicates they are activated (1). These and other simulations can be visualized as animated gif files at <http://sysbiomed-erlangen.weebly.com/resources.html>.

using Boolean modeling (Saez-Rodriguez et al., 2007; Kang et al., 2011). For example, **Figure 3A** is the graphical depiction of a Boolean network representing the triggering of NF-κB signaling, the master controller of the immune response, upon activation of Toll-like receptor 5 (TLR5). This event happens when the bacterial flagellum “is sensed” by the macrophage

upon the binding of the bacterial protein *flagellin* to TLR5 (See Supplementary Material). The network is sequentially organized with the receptor activation as input, the subsequent activation of the NF- κ B signaling pathway at the cell cytoplasm and the triggering of an NF- κ B transcriptional circuit in the nucleus. In the network, nodes account for the network compounds, primarily different biomolecules like proteins and miRNAs, but also the cellular phenotypes triggered by the network. Further, the network edges account for the mutual interactions between the compounds, which are in the model represented like Boolean logic functions.

By combining and integrating this simple logic functions over the network in the course of a computational simulation, one can represent the complex sequential activation of the biochemical network modeled. A computational simulation is the mimicking the behavior of the system in a given biological scenario using the equivalent mathematical model: an *in silico* trigger emulates the correspondent biological signal, and the state of all elements of the model is updated at each iteration step by considering the state they assumed at the previous step, thus imitating the propagation of the signal throughout the network. The simulation considers discrete time points representing the activation state of the network, but the time between two consecutive time points (two steps of the simulation) is always assumed as uniform which does not necessarily reflect the expected biological time. Simulations can be used to predict the behavior of the system in non-tested experimental conditions. For example the Table in **Figure 3C** is a representation of the state of a few key nodes of the TLR5 macrophage network at different iterations in a simulation that mimics the triggering of the system after igniting the input signal (*flagellin* = 1). Blue stands for nodes off (0) at the time iteration considered, while orange indicates they are activated (1).

Examples of Boolean Models in Literature

In an interested case-study, Saez-Rodríguez et al. derived a large-scale Boolean network to represent the activation of T cells (Saez-Rodríguez et al., 2007; Kang et al., 2011). T cells, which belong to the adaptive branch of the immune response, can play a role in the long-term response to lung infection (Chen and Kolls, 2013). The network included the signaling pathways downstream of the T cell receptor, the CD4/CD8 co-receptors, and the accessory signaling receptor CD28. Altogether the network with 94 nodes and 123 interactions includes the primary mechanism behind the activation of T cells and depicts the complexity of biochemical pathways and the reciprocal crosstalk. Saez-Rodríguez et al. exploited one of the main advantages of Boolean networks in their analysis: Boolean models have very low computational requirements for simulation when compared with almost any other model and therefore they scale well with network size (i.e., they can simulate large networks). In line with this, they used their Boolean model to simulate and predict in a systematic and qualitative manner the effect of a large number of gene knockouts ("*in silico knockouts*"). Based on the simulations, the model predicted that antibody-mediated perturbation of CD28 and the genetic knockout of the kinase Fyn, two of the network

compounds, may have relevant effects on the network activation, and these effects could be validated experimentally. Using a strategy similar to that of *in silico knockouts*, Boolean networks have been used to predict the effect of drug combinatory treatments in cancer (Layek et al., 2011). We do not see any formal limitation impeding the use of the same strategy to predict the effect of the combination of antibiotics and immunomodulatory drugs in acute infections like bacterial pneumonia.

In line with this example but in the context of lung infection, Anderson et al. (2016) studied human dendritic cell response against the influenza H1N1, a virus that can co-infect with several types of bacteria to produce pneumonia (Joseph et al., 2013). To this end, they derived a biochemical network with 13 nodes corresponding to genes and transcription factors playing a role in antiviral response (e.g., NF- κ B, STAT1 and IRF1), and 42 edges representing the activation of key immune pathways during the infection. The simulations were done with an asynchronous Boolean model. The initial states of the Boolean simulations were based on experimentally observed expression patterns for the genes in the network (e.g., EGF, NFAT, PDGF and IL-2 set as active during H1N1 virus infection). The model was used to investigate the regulation of the IL-2 pathway after exposure to influenza virus. The model simulations suggested that NFAT can regulate IL-2 signaling in the context of the virus infection, a prediction that was experimentally validated. Further analysis led to the conclusion that IRF and NF- κ B signaling share regulatory functions in H1N1, two out of the three major signaling pathways responsible for mediating TLR-induced responses in viruses, bacteria and other pathogens (Mogensen, 2009).

Although Boolean models are more suited for investigating biochemical networks, they can also be used for describing networks of interacting cell population's exceptions (Jack et al., 2011). For example, Thakar et al. (2007, 2009, 2012) developed a Boolean model for the regulation of the immune system response during the respiratory inflammation caused in mouse by two close relatives of the *Bordetella* genus: *Bordetella bronchiseptica*, a bacterium causing infectious bronchitis in animals, and the human pathogen *B. pertussis*. The model contains well-established knowledge on the immune response after independent infection with each one of the bacterium. The nodes represent (a) immune cell types involved in inflammatory process, including dendritic, T or B cells, (b) cytokines related to a specific phase of the immune response or (c) antibodies. Some edges account for the activation of the immune cells upon stimulation, while others connect the active immune cells to the production and secretion of cytokines and antibodies. Thus, a Boolean network can be used to integrate cell-to-cell and intracellular scale events. In the network, synchronous and asynchronous simulations were performed with the Boolean model. Further experimental data on the host- and pathogen interaction were used to refine the logic gates describing the behavior of the nodes. Model simulations identified three phases in the course of the *B. bronchiseptica* induced inflammation, and suggested that antigen regulatory mechanisms play a prominent role along the whole process, conclusions that were experimentally validated.

In a continued work, the model was expanded by including published experimental data on time evolution for the concentration of IL-10 and IFN γ , information useful to expand the network by including the differentiation of naïve T cells. Model simulations proved to be able to make several predictions namely: (1) the cooperativity between IL-10 and IL-4 signaling to inhibit IFN γ , which was later experimentally validated; (2) the role of the interactions among IL-10, IFN γ , IL-12, and IL-4 signaling in deciding the naïve T cell differentiation process into either Th1 or Th2; and (3) the fact that Th1 cell activity must be temporally longer than that of Th2 cells. To integrate time-series data, the authors transformed the discrete model into a hybrid model (see final section of this paper). Further, the group adapted the network to investigate the co-infection of rabbits with *B. bronchiseptica* and *Trichostrongylus retortaeformis*, a worm that usually infects herbivores inducing a severe infection. Helminth infections predispose mice to pneumococcal pneumonia (Apiwattanakul et al., 2014), and some helminths can trigger pneumonia in humans (Cheepsattayakorn and Cheepsattayakorn, 2014). Using previously published experimental data describing the host immune response to the single infection and for co-infections, an asynchronous Boolean model was derived. Boolean logic functions were derived from literature, and in case of uncertainties, functions were adjusted by comparing the simulation output with experimental results. The resulting Boolean model was used to investigate the crosstalk between regulatory pathways upon the infection with the two pathogens. To validate the co-transfection network, the group infected rabbits with both pathogens, and then assessed the robustness of the model by comparing the resulting activation pattern of the immune response network upon infection with data obtained in rabbit model. Further, simulations representing single knockout of selected network compounds were used to determine central nodes of the single and co-infection networks, with special attention to the knockout of cytokines and immune cell populations' nodes. For example, knockout of nodes accounting for populations of B, dendritic or T cells led to a longer persistence of the bacteria in all case studies. In contrast, knockout for the IL12II or eosinophil population nodes in the co-infection network rendered parasite population not persistent anymore.

Critical Remarks on Boolean Models

There are some alternative modeling frameworks derived from Boolean logic. For example Probabilistic Boolean Networks (PBNs) use Boolean logic and Boolean values, and then implement a set of probabilistic rules determining the state of each node. Each rule is associated to a probability that a specific network state can occur based on the states of its inputs, and the probability for the transition can be assigned based on experimental data (Shmulevich et al., 2002). This probabilistic feature can make PBNs interesting to account for immune cell interactions with a probabilistic compound due to the low abundance of the cells involved at the site of interaction, but also for intracellular interactions with molecules in low abundance (Celli et al., 2012).

Despite the complexity of the interactions that can be modeled by pure or probabilistic Boolean logic, the universe of possible values for every network node is always reduced to 0 and 1. In multi-valued logic models each node can assume several discrete values that refer to a specific qualitative property (for example, “0” for no significant amount, “1” for small amount, and “2” for large amount of receptors activated). This approach has proved to be very valuable in some cases like transcriptional activation. For example, one can have transcriptional targets requiring low levels of active NF- κ B, while others may require much higher levels, and a multi-valued model may be able to account for this distinctive activation pattern. In these models, thresholds can be set to determine the qualitative behavior of the node (Schlatter et al., 2009; Guebel et al., 2012).

Further, Boolean networks can be “calibrated.” This calibration is named pruning and consist on a systematic addition or deleting of nodes or interactions based on the use of quantitative data. In this way, one can make use of -omics data sets to refine the structure of the Boolean network (Terfve et al., 2015). Boolean networks are not suited for spatial features associated to biochemical reactions like molecular gradients. But perhaps the main limitation of Boolean models is their poor ability to reproduce and simulate the non-linearity arising from the existence of regulatory loops in biochemical networks. In consequence, they cannot provide detailed analysis of the fine-tuned regulation of biochemical systems enriched in these motifs. Mathematical models that can handle successfully nonlinearity are those in ordinary differential equations, which are discussed in the context of lung infection and inflammation in the coming section.

MODELS IN ORDINARY DIFFERENTIAL EQUATIONS

Main Features of ODE

Under the assumptions that the biochemical reactions happen in discrete and homogenous intracellular regions (Rahmandad and Sterman, 2008) and the velocities of the biochemical reactions are determined by the concentration of the intervening species (Gustafsson and Sternad, 2013), biochemical networks can be modeled using kinetic models. Kinetic models are a special type of models in ordinary differential equations (ODE), where the equations describe the rate of change of the populations of the biomolecules involved in the biochemical reactions. Similar types of ODE models can be derived to account for the dynamics of interacting cell populations.

To model a biochemical network composed by several molecules, one has to formulate a system of coupled differential equations, consisting on one equation per each element of the system whose dynamics is modeled. For example, **Figure 4** top left is a simplified depiction of an ODE model accounting for two branches of the inflammatory response triggered upon activation of the IL-1 β receptor by its ligand in lung epithelial cells (a detailed scheme can be found in Supplementary Material). One branch mediated by NF- κ B promotes the secretion of several pro-inflammatory cytokines like IL-6, while a second branch

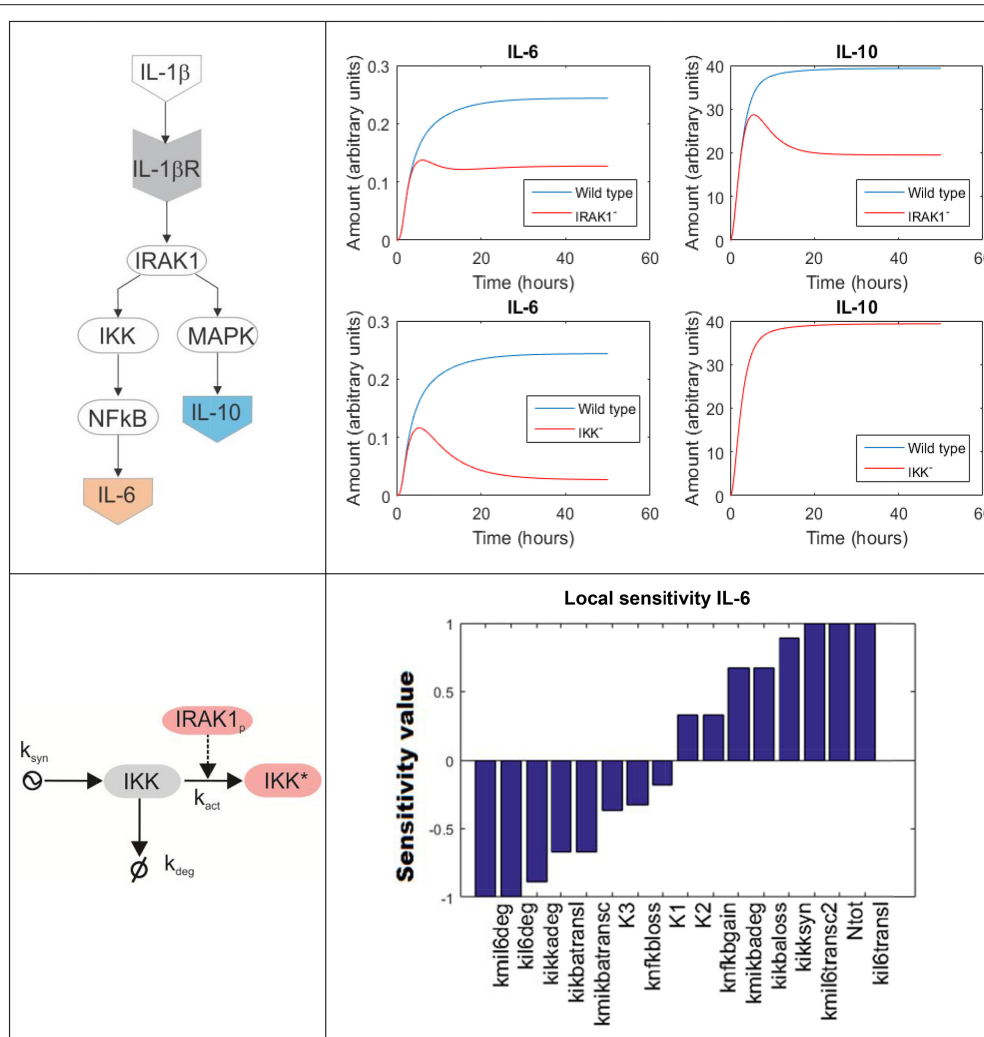


FIGURE 4 | ODE model accounting for the inflammatory response triggered in lung epithelial cells upon activation of the IL-1 β receptor. **Top left:** simplified depiction of the two branches of the inflammatory response triggered upon activation of the IL-1 β receptor included in the model. **Top right:** model simulations accounting for the production of IL-6 and IL-10 in response of IL-1 β mediated NF- κ B activation under IRAK1 (**Top**) and IKK (**Bottom**) knockout conditions. **Bottom left:** detailed scheme of the accounting for the processes affecting in the model the values of inactive IKK. The arrow finishing in null symbol accounts for degradation of IKK, while the arrow starting in a ying-ying symbol stand for synthesis. **Bottom right:** local sensitivities of IL-6 concentration with respect to the perturbation of the model parameters (see Supplementary Material For further details).

mediated by MAPKs promotes the secretion of, among others, the anti-inflammatory cytokine IL-10. The branching point is the activation of IRAK1. The model accounts for the changes on time of the network compounds using the mass-action formalism. For example, the following equation represents the rate of change of the concentration of inactive IKK (IKK, see **Figure 4** bottom left):

$$\frac{dIKK}{dt} = k_{syn} - k_{act} \cdot IKK \cdot IRAK1_p - k_{deg} \cdot IKK$$

In the right-hand side of the equation, each term accounts for a process affecting the concentration of IKK. The first term represents the synthesis of IKK, here modeled like a process at constant ($\text{mM} \cdot \text{h}^{-1}$ units), stable functioning and

represented by the parameter k_{syn} . The second term accounts for the phosphorylation and activation of IKK, represented by a rate equation proportional to the quantities of inactive IKK and phosphorylated IRAK1 (IRAK1_p) and multiplied by a rate constant (k_{act} , $\text{mM}^{-1} \cdot \text{h}^{-1}$ units). The third term models the degradation of inactive IKK, which linearly depends on its concentration and the rate parameter k_{deg} (h^{-1} units).

Contrary to Boolean networks, ODE models can be used to make continuous and precise time-depending simulations. For example, one can simulate the effect of deactivating mutations in key genes of the NF- κ B pathway in the secretion of cytokines by lung epithelial cells. **Figure 4** top right displays a set of predictive simulations accounting for the production of IL-6 and IL-10 in response of IL-1 β mediated NF- κ B activation under knockout

conditions. The wild type condition is displayed in blue. In addition we showed the predicted time profile for both cytokines under deactivating mutations of IRAK1 (here represented as IRAK1^-) and IKK (IKK^-). Compared to Boolean networks, the simulations are continuous, more detailed and give quantitative information about the duration and the intensity of the cytokine secretion in the different conditions simulated. For example, the model simulations indicate that the IRAK1 mutation (IRAK1^-) has a significant effect in the production of IL-10 and can lead to a 50% decrease in its maximal concentration. Similarly, IKK mutation (IKK^-) reduces the secretion of IL-6 while not affecting IL-10. These model predictions match published experiment reports (Supplementary Material).

ODE modeling is a well-established methodology in biomedicine, often included in the training offered in master programs in computer sciences, physics, computational biology and bioinformatics. The key feature of ODE models is the existence of a large array of computational and theoretical techniques of model analysis beyond simple simulations. This includes sensitivity analysis (Savageau, 1971; Zi, 2011; Castillo-Montiel et al., 2015), symbolic analysis (Ibargüen-Mondragón et al., 2014), bifurcation analysis (Duan et al., 2011; Yuri, 2017), design space analysis (Savageau, 2011) model optimization (Vera et al., 2003; Zhang et al., 2015) and parameter estimation and identifiability (Raue et al., 2009).

For example, in sensitivity analysis one can obtain quantitative information on how variation in the value of given model parameters can affect the dynamics and values of the model's time-dependent variables (Saltelli et al., 2000). In our example (Figure 4), we focus on local sensitivities, which are calculated in a narrow region of the model parameter values around the condition of interest, though it is possible to perform sensitivity analysis for a wider interval using global sensitivities (Mathew et al., 2014). In our case, local sensitivity analysis allows for detecting the model parameters affecting the most the maximal value of IL-6 during the simulation, here used as a measure of the production of pro-inflammatory cytokines in the course of cell activation. We computed the local sensitivities by varying the parameter values within a small interval around its value (the value of the parameter set was arbitrarily defined in a way it resulted biologically feasible and instructive for the purpose of this review). The perturbed parameters are ordered in terms of their effect in the maximal value of IL-6, from those whose increase negatively affects IL-6 to those that make a positive effect (Figure 4 bottom right). In a real case-study, the output of this analysis could be used to select promising molecular drug targets for new immune modulatory drugs. These drugs could be administered in parallel to antibiotics and would modulate the production of pro-inflammatory cytokines during the acute phase of the inflammation. A similar approach relying on ODE models and sensitivity analysis has been successfully utilized in anticancer drugs therapy (Schoeberl et al., 2009), and there are no evident limitations to make something similar in bacterial lung infection.

ODE models can account for highly non-linear processes and show properties, often found in biological regulatory circuits, like bistability or oscillation (Tyson et al., 2003). In that

case, advanced model analysis tools can be used with ODE models to dissect the non-linear dynamics of inflammatory and infectious diseases. For example, ODEs can be combined with bifurcation analysis. In bifurcation analysis, advanced methods from non-linear dynamics mathematics are used to detect model parameters associated to key interactions and processes, for which its perturbation in given intervals generate a shift in the equilibrium of the system. Here, we are not talking about smooth, gradual changes like those detected by local sensitivity analysis, but about drastic changes such as those generated by the sudden activation of, for example, the positive feedback circuits behind several known autocrine loops in inflammation (Coward et al., 2002). Dunster et al. (2014) employed bifurcation analysis of an ODE model to analyze the role of different immune cells on the resolution of inflammation. The model accounted for the interactions between macrophages, neutrophils and pro-inflammatory mediators like $\text{T}\alpha$ and IL-8. The model analysis focused on finding the bistability region in the model, that is, the set of model configurations in which the system can switch between the two physiological states: inflammation and resting. Based on their analysis, they concluded that key processes accelerating the resolution of inflammation are an increase of macrophage phagocytosis and the neutrophil apoptosis.

Moreover, these ODE model analysis methods can be integrated in workflows to investigate complex properties of biological systems (Nikolov et al., 2010). A very recent work created and analyzed a mathematical model of the *Streptococcus pneumoniae* lung infection (Domínguez-Hüttinger et al., 2017). It includes the interactions between the pathogen and the host like macrophages and neutrophils activation, bacteria clearance, epithelial cell barrier integrity and bacteria migration through the barrier to the vessels. In the model, the authors differentiated between a commensal state, that does not produce a disease, and an invasive and infective state of the bacteria. By including this feature in the bacteria population dynamics, the model predicted four different possible phenotypes: (i) sepsis, that is systemic bacteria spread and inflammation, (ii) immunological scarring, that is, cumulative, long-lasting immune response to pathogens inducing tissue remodeling and altered immune responses to new pathogenic challenges; (iii) sepsis + immunological scarring, or (iv) healthy infection recovery. Further, model simulations were used to assess the required duration of antibiotic treatment to treat each phenotype.

Sparse information taken from the literature can be used to characterize model parameters. Based on predictive model simulations using the data-based model, one can gain new insights on the regulation of the network underlying, for example, pathogen associated tissue destruction. The immune system residing in the respiratory mucosa has to achieve a balance between its ability to deplete pathogens and to induce tissue damage; a failure in this tightly control mechanism can induce chronic inflammation and tissue destruction (Lugade et al., 2011). Lo and coworkers (Lo et al., 2013) constructed and characterized with available data a model accounting for the abnormal regulation of T helper 1 (Th1), T cell helper 2 (Th2), and T regulatory cells (Treg) in chronic lung mucosal inflammation. The model was used to simulate possible

physiological scenarios concerning inflammation of the lung mucosa. Based on the model simulations, the authors found that deregulation of the interaction between these immune cells is sufficient to explain the emergence of chronic lung mucosa inflammation. Specifically, the model predicts that upon Treg downregulation the Th1 and Th2 responses to cytokine can be abnormally high. Since it is known that mucosal Th1 and Th2 cells can produce pro-inflammatory cytokines (Neurath et al., 2002), the system displays the structure of an autocrine positive feedback loop, which could induce under deregulation signal amplification and chronic inflammation.

In predisposed patients, airways and lung infections caused by both viruses and bacteria can unbalance the regulation of the local lung immune system and contribute to asthma exacerbation (Pelaia et al., 2006). Chernyavsky et al. (2014) derived an ODE model on the emergence of airway smooth muscle cells (ASMC) hyperplasia due to asthma-related inflammation, which was characterized using published data from biopsies and inflammatory biomarkers (Contoli et al., 2010). The authors modeled interactions between proliferative and non-proliferative ASMCs and their impact on the inflammatory state of the lung. The model was utilized to simulate the development of the asthma associated inflammation. Model simulations showed that the speed of inflammation resolution is a leading factor in the long-term evolution of asthma, and also that the features of the tissue remodeling during and after the inflammation are important to control the long-term evolution of asthma.

The parameters of ODE models can be estimated by fitting the model simulations to dense time series of experimental data in a process called model calibration. For example, Mochan et al. (2014) modeled pneumococcal lung infection and used time series data from titrated mice infection to calibrate the model. Bacteria titration refers to the inoculation of different initial amounts of bacteria to mice. The model included the interplay between the bacteria, lung epithelial cells and alveolar macrophages, the production of cytokines and chemokines and the subsequent recruitment and activation of neutrophils and monocytes. The model was used to simulate and quantify the dynamics of the damage in the tissue caused by the immune system in the early phases of infection. The model simulation analysis pointed to the importance of the dynamics of macrophage phagocytosis to explain the differences between the phenotypes of resistance or sensitivity to pathogen. In a different work, Guo et al. (2011) integrated time series data of bacterial burden in an ODE model to quantify the contribution of neutrophils on the bacterial clearance during pneumonia in mice. To this end, the authors formulated a single-equation model accounting for the dynamics of bacterial growth when exposed to lung neutrophils. The model not only correctly predicted the number of neutrophils that is necessary for suppressing *A. baumannii* growth by 50%, but it also proved to be able to make predictions for the case of infection with other pathogens like *P. aeruginosa*.

Examples of ODE Models in Literature

Smith et al. (2011) built a model accounting for the role of resident alveolar macrophages, neutrophils and

monocyte-derived macrophages in early lung infection by *S. pneumoniae* in mice. The model includes time-dependent variables for the bacteria population, resting and active macrophages, activated and non-activated epithelial cells, cytokines, neutrophils and the debris associated to infection and tissue damage. To assign value to the model parameters they extracted information from literature, but also fitted their model to time series data for different bacterial titration. The model was used to quantify the contributions of cytotoxicity and immune-mediated damage in pneumococcal pathogenesis. When the authors generated two alternative versions of the model with or without monocyte-derived macrophages recruitment, the dynamics of bacteria growth was not affected. Based on the previous work, Schirm and coworkers proposed a modified mathematical model of cellular interactions in bacterial pneumonia (Schirm et al., 2016). They considered in the model alveolar macrophages, neutrophils and monocyte-derived macrophages. This model was fitted with large time-series data sets from infected mice, which includes measurements for pneumococci, neutrophils and macrophage populations, as well as for IL-6 and debris, here assimilated to histological damage score measured in the lung tissue. The calibrated model was used to simulate the evolution of the disease with or without antibiotics treatment. To this end, the model simulated the administration of 0.02 mg/g Ampicillin or 0.1 mg/g Moxifloxacin every 12 h, starting 24 h after infection. The model simulations indicate that alveolar macrophages are responsible for the quick elimination of the disease. Moreover, the model simulations predicted that the remission of the infection can happen with lower doses of antibiotics than those applied in the experiment. In line with this, the authors propose to utilize model simulations to design alternative time schedules for the antibiotic treatment. This strategy could be relevant in the context of bacterial infection induced sepsis. Sepsis is a common cause of acute kidney injury and therefore a modeling-based methodology for accurate antibiotics dosing could be relevant for critically ill patients (Eyler et al., 2011). To this end, one can derive a pharmacokinetics and pharmacodynamics ODE model accounting for the toxicity and effectiveness of antibiotics, similar to existing models accounting for efficacy vs. toxicity of anticancer drugs (Ballesta et al., 2011).

Coinfections, the co-occurrence and potential synergy between two infectious agents, have been also investigated with ODE models. An example of modeling coinfection is the work by Smith et al. (2013), in which coinfection of mice with influenza virus and *S. pneumoniae* in the lung was investigated. The model includes variables accounting for the dynamics of influenza virus, *S. pneumoniae*, alveolar macrophages and influenza lung epithelial target cells. The model was calibrated using time-series data for the amount of bacteria and virus. Remarkably, the model simulations showed the rebounding in the populations of the bacteria and the virus. Pathogen rebounding is the proliferation of a pathogen after an initial decrease when it co-occurs with a second pathogen. In Smith et al (Domínguez-Hüttinger et al., 2017), upon infection with bacteria, the virus population rebounds due to the release of viruses that were latent in the immune and lung cells killed by the bacteria. In parallel, the

model predicts an increase in the bacterial load due to the impairment of macrophage response provoked by the presence of the viruses. The system as described displays the structure of a positive feedback loop in which bacterial and virus infection amplify each other.

Critical Remarks on ODE Models

ODE models can account for important spatial features like molecular gradients only in a very limited manner. An extension of ODE models in this regard could be partial differential equations (PDE) models; however, the lack of appropriate experimental data for their characterization has limited their development in biology to a few but promising case studies (Matzavinos et al., 2004; Murano et al., 2014). In an ideal setup, ODE models require numerous and rich time series data sets for model calibration, a prerequisite to obtain a trustable model. This necessity for complex data sets is a clear limitation, especially when trying to model large biochemical networks. A fundamental limitation of ODE crucial for some biological systems and transcriptional circuits is that predictions based on ODE models may fail for systems with low copy numbers for the molecules or the cells involved in the interactions, in which randomness in their dynamical behavior emerges. These special features are better represented by stochastic models, which are discussed in the coming section.

STOCHASTIC MODELS

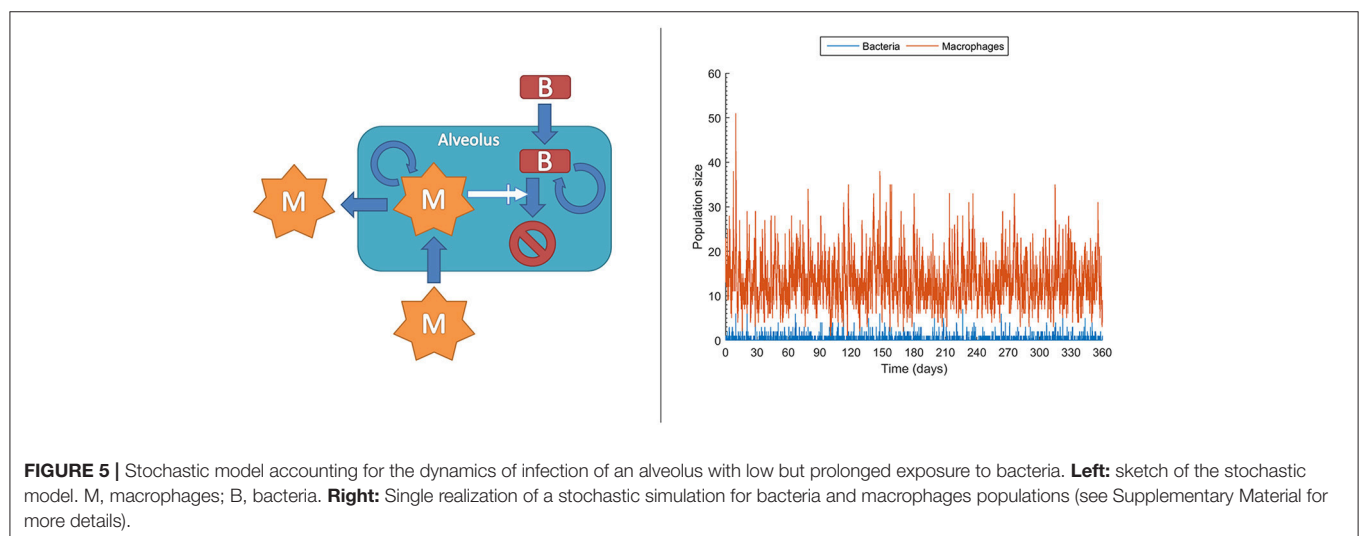
Main Features of Stochastic Models

At the molecular level, chemical events, including biochemical reactions, occur randomly. Taking this strong assumption, it is impossible to deterministically predict when the next reaction occurs, but also each experimental repetition of a biochemical reaction will intrinsically differ in the measured values. This effect is actually important under low copy numbers for the molecules intervening in the reaction, conditions under which it is known and it has been experimentally confirmed

that accuracy collapses for deterministic models like those in ODEs. In contrast, stochastic models can account for this effect rather than attributing it to measurement errors, thereby outperforming deterministic models (Gillespie, 1992; Klipp et al., 2009; Pahle, 2009; Wilkinson, 2009; Ullah and Wolkenhauer, 2010). In stochastic models, chemical species or cell populations are represented as discrete random variables. These variables form the state space of the stochastic model and describe the abundance of each species at any given time point. Chemical reactions or cell interactions are envisioned as random processes that change the abundance of the involved species. While these reactions occur randomly, their probability of occurrence depends on the current state and it changes as the system moves from state to state. For example, in the very early phases, both bacteria and macrophages display very low copy numbers, sometimes with single macrophages patrolling one or more alveolus. In these conditions, even small random fluctuations can have a large impact on the population dynamics and therefore a stochastic model is an option for describing their population dynamics. **Figure 5** left displays the structure for a stochastic model, adapted from Van Furth (2012), accounting for the long time dynamics of infection of an alveolus exposed to stochastic bacteria colonization. In the model, the current number of macrophages and bacteria is denoted by m and b respectively. The interactions between macrophages and bacteria determine the state transitions, that is, the increase or decrease of the bacteria and macrophage populations. For example, the stochastic model accounts for the generation of a macrophage (a_{M+}) with the following equation:

$$a_{M+}(m, b) = c_{Mmigrate} + c_{Mbirth} * m + c_{Mresponse} * b * m \quad (1)$$

Here, it is assumed that the generation of a macrophage can occur in three different ways: (i) macrophage migration into an alveolus occurring at a constant probability rate ($c_{Mmigrate}$); and (iii) recruitment of additional macrophages depending on the current number of bacteria and macrophages ($c_{Mresponse} * b * m$). **Figure 5** right is a single long time simulation of the model. The single



simulation reveals large variability in the populations of bacteria and macrophages. In particular, the macrophage population shows large fluctuations, with values ranging from one to up to 51 macrophages in the alveolus, in conditions with very small amount of bacteria. When one performs a large amount of similar simulations (here 10^4 simulations) one can verify that these fluctuations render the fate of the system stochastic. Thus, in a small fraction of the simulations (0.1%) the population of bacteria gets higher than 100. The stochastic model simulations suggest that, under healthy conditions and for low long term lung alveolus exposure to bacteria, most of the episodes of bacteria colonization are quickly resolved, although there is still a small probability of bacterial infection.

As discussed before, the core regulatory pathway controlling the activation after bacterial lung infection of epithelial and immune cells is the NF- κ B pathway. There are two features that make stochastic modeling suitable for investigating NF- κ B activation. Stochastic models are especially suited for transcriptional circuits because gene expression is widely considered to be a process dominated by randomness (Elowitz et al., 2002; Kaern et al., 2005; Wilkinson, 2009; Bressloff, 2017). NF- κ B is a transcription factor, and under some conditions the pathway activation may lead to a low amount of transcriptionally active NF- κ B molecules. In this case, large fluctuations may appear in the transcription of NF- κ B targets, making advisable the use of stochastic modeling. In line with this and using a microfluidic cell culture platform and single cells resolution, Tay and collaborators investigated the features of NF- κ B activation for a wide range of values of concentration for TNF α , one of the infection-associated ligands promoting NF- κ B activation. Under low TNF α concentration, they found single cell heterogeneity and digital response of the cells. This translates into an all-or-none activation pattern for 3–50% of the cells at concentrations as low as 0.1–0.01 ng/ml. To elucidate the regulatory features inducing this behavior, the authors derived a stochastic model accounting for the NF- κ B activation. Using the model, they found that the ability of the model to reproduce the digital response observed relied in the inclusion in the model equations of specific features of TNF α ligand and receptor turnover. Precisely, they found it was related to the limited TNF α amount present in the microfluidic chambers, the TNF α degradation and turnover and the cell-to-cell variability in the amount of TNF α receptor available for activation. Further, to reproduce the data the model assumed a non-linear nature to the IKK activation profile, attributed to the fact that IKK subunits IKK- α and IKK- β achieved full activity when phosphorylated at two different residues (Tay et al., 2010).

In addition, stochastic models are suitable for assessing the fine regulation of feedback loop circuits displaying oscillations or bistability because stochastic models can assess their sensitivity to small random perturbations (Levine et al., 2013; Dobrzyński et al., 2014). NF- κ B signaling is controlled by a combination of intracellular negative feedback loops, which are able to induce oscillations (Nelson et al., 2004), and autocrine positive feedback loops with the ability to trigger bistable switches (Pękalski et al., 2013). In both cases, stochastic modeling is the right tool for assessing the sensitivity of NF- κ B signaling to small random

perturbations induced by these regulatory loops. Ashall et al. combined single-cell life imaging and modeling to investigate the role of these oscillations. They could show that the expression of a number of NF- κ B transcriptional targets depends on the frequency of the potentially pulsatile inflammatory signals found at the site of inflammation and infection. Although these features could be investigated by ODE modeling, the heterogeneity of single-cell responses they found exceeded the capabilities of these models. However, a stochastic model that assumed delayed stochastic transcription for I κ B α and stochastic transcription of I κ B α and A20 (all of them inhibitors of NF- κ B signaling embedded in negative feedback loops) proved to be able to recapitulate the cell-to-cell heterogeneity in the NF- κ B oscillations. In line with these results, the same team recently showed the existence of single cell NF- κ B-mediated oscillatory responses even under physiological concentrations of TNF α , a cytokine that play a pivotal role in the pathogenesis of pneumococcal pneumonia (Takashima et al., 1997; Ashall et al., 2009; Turner et al., 2010).

Other immune related intracellular pathways may display the features that make necessary the use of stochastic modeling. For example, intra- and extra-cellular calcium signaling plays an important role in the immune response (Vig and Kinet, 2009) and they have been described using stochastic models (Rüdiger, 2014). Further, TRAIL-mediated apoptosis, a mechanism playing a role in limiting the effect of alveolar macrophages on the extension of inflammation during *S. pneumoniae* lung infection (Steinwede et al., 2012), can display stochastic cell-to-cell variability in its activation (Bertaux et al., 2014). The dynamics of pathogenic bacteria intracellular circuits can become also stochastic (Norman et al., 2015). In line with this, Tuchscherer et al. (2011) showed that as part of their immune escape strategies, *Staphylococcus aureus* can induce a phenotype switching. Bacteria switching is a transient bacteria phenotypic change, governed by intrinsic stochasticity intracellular circuits, that provides bacteria with functional diversity and fast adaptation to environmental changes.

Examples of Stochastic Models in Literature

Stochastic models have been used for decades to dissect the cell population dynamics during lung infection. Two recent papers deal with the lung infection by *Francisella tularensis* (Gillard et al., 2014; Wood et al., 2014), an infectious intracellular gram-negative bacterium that infects primarily macrophages. When inhaled in an aerosol, *F. tularensis* can proliferate in the lung causing a type of severe pneumonia called pneumonic tularemia. Gillard et al. (2014) derived a stochastic mathematical model accounting for the early phases of *F. tularensis* pathogenesis in the lung. The model contained three possible states for the alveolar macrophages, coinciding with three of its most prominent phenotypes: (1) resting macrophages, functional but with no ability to kill bacteria; (2) suppressed macrophages, unable to overcome cytokine production and bacteria phagocytosis; and (3) classically activated macrophages, which play a role in clearing the infection. Regarding the dynamics of macrophages,

the model considers as key events in the early infection phase the macrophage infection, its suppression and activation and death. Concerning the bacteria dynamics, the model accounts for bacterial proliferation, death and phagosome escape to the cytosol. To derive the model, the authors extended the framework of the **birth-and-death processes stochastic models** by attributing to each macrophage four features (spatial location, state of activation, number of phagosome bacteria, number of cytosolic bacteria) and making them affect the macrophage and bacteria populations dynamics (Levy and Green, 1968; Tranquillo et al., 1989). The model was able to reproduce most of the knowledge available on the early phases of the *F. tularensis* infection, but the authors claimed it could further provide insights into potential coadjuvants of antibiotic therapies, aiming at stimulating macrophage activation. Finally, since it exceeds the scope of this review, we do not discuss here but want to mention the use of stochastic modeling in the simulation and prediction of epidemics spread of bacteria-associated lung infection diseases (Grundmann and Hellriegel, 2006; D'Agata et al., 2007; Agliari et al., 2013).

Critical Remarks on Stochastic Models

Stochastic models do not scale well with the size of biochemical networks due to their structural complexity and the necessity to perform multiple realizations of the same simulation. However, the exponential increase in the computational power will make possible in the close future to simulate large stochastic models even in average scientific workstations. Calibration of stochastic models requires high sensitivity and specificity experimental techniques capable of quantifying random effects and fluctuations in molecule or cell abundance. For biochemical systems, this translates into single-cell technologies like single-cell transcriptomics, single-cell PCR, mass cytometry and fluorescence-based technologies (Crépieux et al., 1997; Lidke and Wilson, 2009; Spiller et al., 2010; Bakstad et al., 2012; Bendall and Nolan, 2012; Haack et al., 2013). Although these methods are to date technically challenging, expensive and not available in an average cell biology lab, one can foresee they will become standard technologies in relatively short time. Altogether, stochastic models are currently not suitable for systems that include many different interacting molecular or cellular species.

AGENT BASED MODELS

Main Features of ABMs

Many if not most of the intracellular biochemical reactions happen in complex, often highly crowded and heterogeneous spatial compartments (Rivas et al., 2004; Minton, 2006). Similarly, cell-to-cell interactions are affected by the features of the tissue compartments in which they take place. Logic networks, ODE or stochastic models have a relatively limited ability to account for spatial features. In contrast agent-based models (ABM) are powerful tools to simulate in a detailed manner the spatial features of these interactions at the single molecule or cell level. Agent-based models can be used to simulate the dynamics of ensembles of so-called agents in two

and three dimensions predefined spaces. Agents are entities mimicking molecules or cells, which have the ability to simulate their movement within the modeled space compartment and their interactions with other species, also modeled like agents. The fate and movement of the agents depends on a set of rules, which are based on their molecular and cellular properties and the features of their interactions. ABMs can include a variety of different agent populations, which could operate at different spatial scales within the model. The **environment** surrounding the agents can display multiple spatial heterogeneous features, like spatial domains with different ability to diffuse or interact for the agents. Finally, the **rules** defining the update of the agent behavior can be the result of other models like ODEs or Boolean networks, but also stochastic rules. Ultimately, agent-based model simulations are intended to find collective, emergent patterns in the behavior of the agent populations. In the biomedical context ABMs have been primarily used to investigate interactions between cell populations. For example, in the early phases of infection both bacteria and macrophages are in low numbers and the spatial aspects of macrophage motility, sensing and recruitment, or bacteria motility and proliferation may decide the conditions for a fast resolution or a long-lasting extended infection. In these conditions, ABMs offer the possibility to simulate with detail the spatial features of the interaction between macrophages and bacteria in the lung alveolus. **Figure 6** accounts for simulations made with an ABM. The ABM stands for the dynamics of two populations of agents accounting for bacteria and macrophages at the very early phases of bacterial lung infection. Thus, the infection is assumed to take place in a single alveolus and both agents are assumed in low numbers when the simulations are initiated. The alveolus is modeled like a torus shaped surface of 32×32 pixels. The macrophages are 2 pixels wide and bacteria are considered to be non-dimensional dots. During the simulations, bacteria and macrophages move in 1 pixel. In the simulations, the time is discrete, with time iterations in the time-scale of the processes considered. As initial conditions for the simulations, the initial amount of bacteria and macrophages are situated in random positions of the 2D space. The behavior of each individual agent is governed by a set of rules describing the ability of macrophages and bacteria to move, the bacteria proliferation, the recruitment of monocyte-derived macrophages and the bacteria killing after bacteria-macrophage encounter (See Supplementary Material for more details). To make the model more accurate, we assumed the stochasticity for the bacteria movement and proliferation, as well as for the macrophage movement and recruitment. Thus, the evolution and final fate of two similar simulations can differ drastically. For example, **Figure 6** top displays the time course for bacteria and macrophage populations during two similarly initiated simulations with 250 time units duration, which display totally different time courses. In the top simulation, the bacteria infection is resolved and the bacteria population gets extinct, while the bottom simulation ended with a successful bacterial colonization although the initial conditions were very similar.

In many ABMs like in this one, a number of the processes models are described by stochastic rules. Thus, the simulations become stochastic and to detect patterns of regulation ensembles

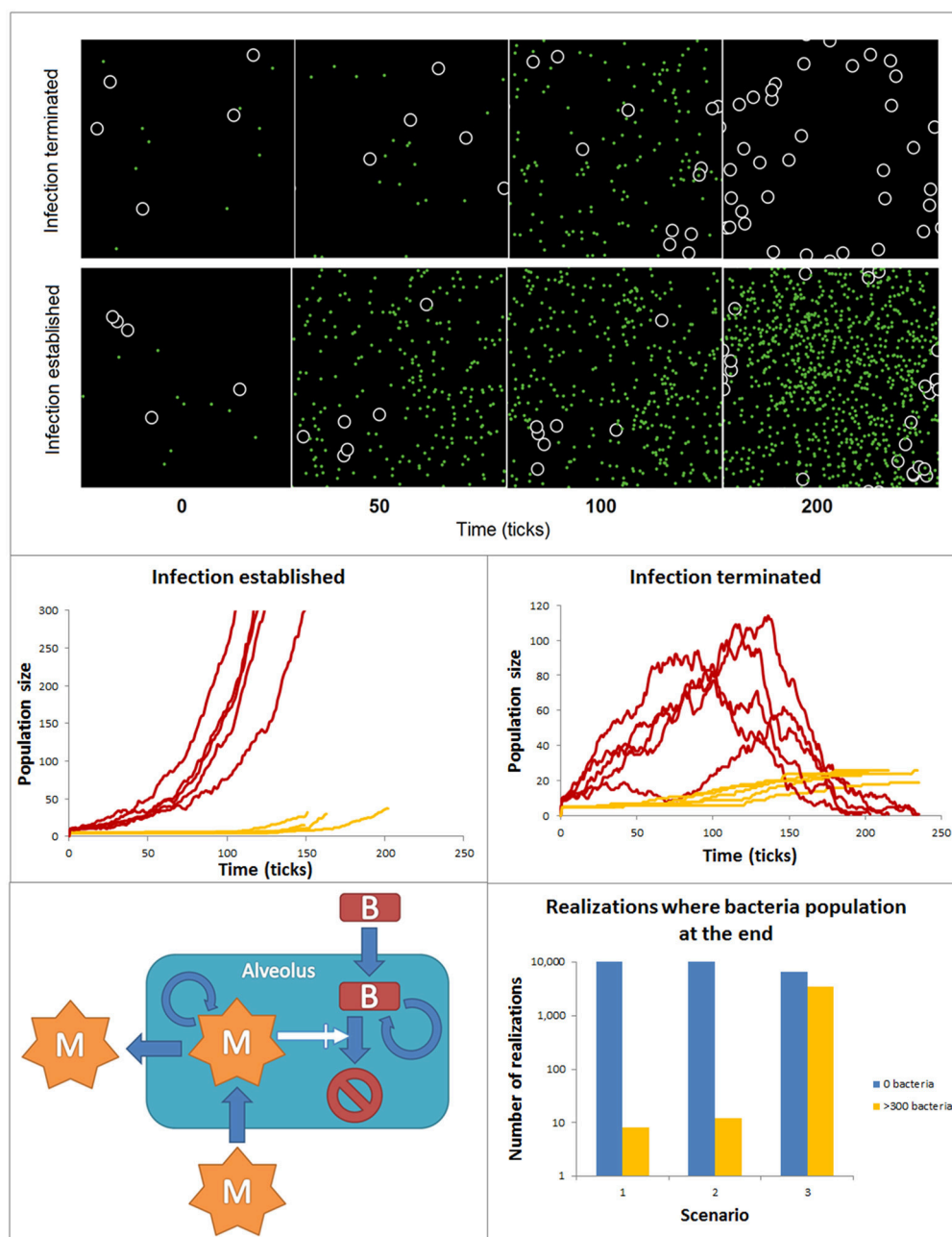


FIGURE 6 | ABM accounting for the spatial features of bacteria and macrophage dynamics during early phases of lung alveolus infection. **Top:** Time course for bacteria and macrophage populations predicted by the ABM for simulations similarly initiated but reproducing infection resolution (top) and infection establishment (bottom). The black space represents an alveolus, white circles are macrophages and green dots are bacteria. Both populations can freely move within the alveolus. **Center:** 10 similarly initiated ABM simulations are classified into those accounting for infection successfully established (left) and those representing infection resolved (right). Red lines represent bacteria populations and yellow lines macrophage population. **Bottom:** ensembles of ABM simulations used to assess the relative importance of the processes modeled in the simulation output. 10^4 simulations were implemented for each scenario. The solutions were divided into two groups (a) yellow bar: solutions ending in an establishment of bacterial infection [bacteria (t_i) ≥ 300]; and (b) blue bar: solutions ending with depletion of the bacteria population (bacteria (t_{final}) = 0).

of ABM simulations are analyzed using statistical methods. In our example, we performed a series of simulations and classify them in two groups of 5 simulations (**Figure 6** center): (a) those in which at the end of the simulation the population of bacteria is extinguished and (b) those in which the population of bacteria reaches 300 individual in the course of the simulation, used as

indicator that the bacteria colonization has been established and the infection has extended to surrounding alveoli. In line with this, ensembles of predictive simulations can be used to assess the relative importance of the processes modeled in the simulation output. For example, we used the model to assess the effect on the success of bacteria colonization of higher proliferation rate

of bacteria invasion (scenario 2) and decreasing infiltration of macrophage (scenario 3, **Figure 6** bottom). Scenario 1 defines the control situation. To make this analysis, we run 10^4 ABM simulations for each scenario, and counted the number of simulations per scenario in which the bacteria population was extinguished (blue bar) or the bacteria colonization was successful (orange bar). The results show a certain level of stochasticity and suggest that decreased efficiency in monocyte-derived macrophage recruitment has more impact in fostering bacteria colonization than increased bacterial proliferation rate.

Examples of ABMs in Literature

Chavali et al. made a detailed discussion of the use of ABMs to investigate and characterize emergent properties of immunological systems (Chavali et al., 2008). ABMs have been used to model in detail the spatial features of molecular interactions within cellular compartments, for example, the dynamics of molecules in cell membranes (Haack et al., 2013; Santos et al., 2016). In line with this, Rhodes et al. employed agent-based modeling to analyse the spatial features of the cytoplasmic dynamics for the NF- κ B inhibitor I κ B α (Rhodes et al., 2015). It has been found that I κ B α can co-localize and get sequestered in cytoskeleton structures like the microtubule organizing center and the α -tubulin filaments (Crépieux et al., 1997). To model in detail this process, Rhodes and co-workers derived a model for the NF- κ B activation via type 1 IL-1 receptor (IL-1R1). The model considers: (1) activation of NF- κ B through IL1R; (2) activation of anti-apoptotic pathways via PI3k signaling; and (3) cytoskeleton reorganization during the NF- κ B activation through Ras activation. Using model simulations, the authors hypothesized that the sequestration of I κ B α can be a mechanism to modulate the intensity of the L1RI input signal coming from L1RI when transduced inside the cell. The mobilization and/or sequestering of signaling proteins to microtubules and other cytoskeleton structures has been found in other key pathways for inflammation like MAPK cascades (Hanson et al., 2007), which indicates that the use of ABMs to dissect the fine-tuning of this mechanism may render interesting mechanistic hypothesis.

ABMS can also be used to establish the link between molecular interactions and cell phenotypes. In line with this, Stern et al. used an ABM to simulate the response to damaged tissue and barrier disruption signals of individual epithelial cells embedded in an extracellular matrix (Stern et al., 2012). In many infectious diseases including pneumonia, the breakdown of the epithelial barrier exposes the inner part of the organism to external pathogens and facilitates their systemic spread and the emergence of sepsis. In the model used, the agents account for the epithelial cells and the rules for the effect on them of the activation of the EGF and TGF- β receptor mediated signaling pathways. It has been found that down-regulation of TNF- α signaling and activation of EGFR signaling contribute to the maintenance of epithelial barrier integrity and function in lung and other epithelial tissues (Finigan et al., 2012; Patel et al., 2013; Uwada et al., 2017). The model was able to simulate tissue damage and wound recovery. Moreover, the model simulations suggested the existence of a mechanism for the crosstalk between TGF- β and EGFR pathways involved in the recovery after damage. The activation of these pathways have been linked to the response

alveolar epithelial cells to some types of bacterial infection (Choi et al., 2011; Li et al., 2015).

ABMS can also be utilized to dissect the spatial features of cell-to-cell interactions in their natural tissue compartments. In order to investigate T cell (TC) activation Bogle and Dunbar built an ABM (Bogle and Dunbar, 2010). The model attempted to investigate the spatial features of TC activation by active dendritic cells (DCs) in the lymph node, thereby trying to establish mechanistic links between the properties of TC and DC motility in the lymph node and the timing and strength of the TC response elicited. The processes included in the ABM were the proliferation of TCs in lymph nodes, the DC driven activation of lymphocytes, and the DC and TC trafficking through the lymph node. The model was used to simulate the proliferation, release and changes in the affinity profile of TCs in the lymph node. The simulation results correlate with data accounting for the efflux rate of activated TCs from lymph nodes. Further, model analysis and simulation were used by the authors to point to open questions and gaps in the current knowledge of the TC-DC interaction in lymph nodes. For example, they hypothesized that the deeper understanding of TC activation can benefit from experiments elucidating the dynamics of the lymph node vascularization, a process that seems to be modulated by the DCs (Webster et al., 2006).

Moreover, ABMs can be used to study in detail spatial properties of infection-related autocrine and paracrine loops. In a work on chronic asthma, a condition we already linked to lung infection, Pothen et al. (2015) hypothesized that in healthy individuals antigenic stimulation drives both the onset and the recovery after allergic inflammation. Under these conditions, allergic inflammation can become a self-limited event. Based on this idea, Pothen et al. used modeling to investigate under which conditions a failure in this process can provoke the chronic airway inflammation associated to asthma. To this end, they derived an ABM that considers spatial features of the interactions between pro- and anti-inflammatory cells during tissue damage and repair in unresolved allergic inflammation. Models simulations suggested that the ability to recover after the allergic episode is in general terms very robust regarding most of the pro- and anti-inflammatory cells interactions, but appears very sensitive to increase in the recruitment and activation of pro-inflammatory cells like neutrophils and eosinophils. The model simulations indicated that down-modulation of pro-inflammatory cell activation could be a therapeutic strategy against the allergic inflammation.

ABMs can be used to mimic the effect of cell exposure to diffused extracellular ligands, biomolecules and non-organic particles. Brown et al. used an ABM to investigate lung inflammation and fibrosis following particulate exposure (Brown et al., 2011), an environmental condition that can increase the chances and severity of lung infection (Mehta et al., 2013). The model accounted for the interaction between lung macrophages and fibroblasts through TNF α and TNF β . It also considered the tissue damage caused by TNF α and the production of collagen for repairing the tissue. The model simulations predicted three main states for particulate exposure associated lung inflammation: (1) self-resolving inflammation, (2) localized tissue damage and fibrosis and (3) elevated pro and anti-inflammatory cytokines

and persistent damage. Model simulations showed that the switch between the different states depends on the intensity and duration of the exposure to the particulate damage.

Critical Remarks on ABMs

ABMs can deal with systems that are complex and heterogeneous from a spatial perspective, but also with biological systems involving many different interacting entities, cell and/or molecules, and multi-levels. The essentially modular structure of ABMs facilitates the addition of new types of agents, accounting for new cellular or molecular players. Even simple rules defining the interactions between the agents can generate extremely complex spatio-temporal regulatory patterns. However, to date these models do not scale well with respect to the number of total interacting agents due to the large computational resources necessary to simulate systems with large number of agents. In line with this, a lot of work has been done in the last decade in terms of methods for efficient and distributed ABM simulation (Aaby et al., 2010). Further ABMs are suited for performing detailed simulations, but very poor in terms of analytical tools. Far from the much elaborated algorithms conceived for the calibration of ODE and PDE models, very little has been done in terms of the systematic integration of quantitative data into ABMs (Bianchi et al., 2007) and computational tools specially designed for modeling of biological systems (Kang et al., 2014; Starruß et al., 2014). In any case, we think that ABMs will be an interesting alternative in the coming future for modeling bacterial lung infection.

DISCUSSION

Great Expectations for Mathematical Modeling in Lung Infection and Inflammation?

We have great expectations in terms of what mathematical modeling can contribute in the coming decade to the

understanding of lung infection pathophysiology. In the last years modeling has been used in biomedicine essentially for integrating multiple types of experimental data, formulating mechanistic hypotheses, or in performing simulation-based therapy assessment. However, mathematical modeling can be used in many other avenues that are not yet sufficiently tested in pulmonology. Epstein (2008) suggested up to 16 motivations other than pure prediction to use modeling and simulation in science. In **Table 1** we have selected a few of them and elaborate how they could be implemented in the context of bacterial lung infection.

To mention an interesting open question, some immune cell types have a dual, often ambiguous role during infection. For example, macrophages and neutrophils are major players in the quick resolution of infection, but under exacerbation they can also worsen the condition by promoting tissue destruction or overwhelming inflammation (Nouailles et al., 2014). This duality can be explained at least in part by the deregulation of intra- and inter-cellular positive feedback loops working often in an autocrine or paracrine manner. For example, TNF α can be secreted by activated macrophages to signal other immune cells in early lung infection (Mukhopadhyay et al., 2006), but it can at the same time promote activation of resident or monocyte-derived macrophages in a amplification loop that can exacerbate local inflammation (Gane et al., 2016). The use on mathematical models dissecting the structure and fine regulation of these circuits can contribute to the understanding of this aspect of acute lung infection.

Moreover, a number of infections and inflammatory conditions in the lung like asthma and tuberculosis persist despite treatment and reappear in an episodic or cyclic fashion. This suggests that autocrine and paracrine regulatory circuits, including positive and negative feedback loops may get disrupted and deregulated in the course of these diseases. For example, G-protein-coupled adenosine receptors have been associated

TABLE 1 | Ten “not-yet-considered” motivations to use mathematical modeling in bacterial lung infections.

| Motivation | Usage in bacterial lung infection | Article of example |
|--|--|-----------------------|
| Discover new questions | Combination of model derivation, simulation and analysis used to formulate in a consistent manner a new hypothesis on the early steps of bacteria lung infection mechanism | Gillard et al., 2014 |
| Guide data collection | Model simulations used to (a) decide the design of the which most suited experiments to test the above hypothesis and (b) select relevant public available data | Thakar et al., 2007 |
| Explain (very distinct from predict) | Customized model simulations used to (a) illustrate the experimental results and (b) discuss/extrapolate the consequences of the proved or disproved hypothesis | Smith et al., 2011 |
| Illuminate core dynamics | A model comprising the core of the network controlling inflammation used to point the key molecules and processes controlling it | Krishna et al., 2006 |
| Reveal the apparently simple to be complex | The analysis of a model representing the apparently simple and small network controlling early bacterial lung infection used to suggest the existence of non-linear behavior associated to feedback loops circuits | Nikolov et al., 2010 |
| Reveal the apparently complex to be simple | Model reductions techniques on a large network representing bacterial lung infection applied to detect the few key processes and molecules controlling the process | Guo et al., 2011 |
| Expose prevailing wisdom as incompatible with available data | Simulations of a mathematical model encoding the current knowledge on molecular interactions controlling initiation of inflammation employed to show inconsistencies with new data | Hoffmann et al., 2002 |
| Bound outcomes to plausible ranges | Comparison between model simulations and available data used to establish the interval of biologically feasible features (parameters) for bacteria proliferation and spread in the lung alveoli | Mochan et al., 2014 |
| Offer crisis options in near-real time | For a patient entering the Intensive Care, personalized model simulations used to predict the course of the host-pathogen interactions and near-real time decide on the therapeutic alternatives | Dix et al., 2016 |

to protection from tissue damage in infection and sepsis (Csóka et al., 2010). Further, adenosine has been linked to the pathogenesis of asthma (Brown et al., 2008). This role is mediated via a physiological negative-feedback mechanism that seems to participate in limiting and terminating tissue-specific and systemic inflammatory responses (Ohta and Sitkovsky, 2001). Mechanistic mathematical modeling of this type of paracrine feedback circuits may shed light into their role controlling overwhelming immune response and the consequences of their deregulation.

Modeling has been used for longtime in pharmacology to assess the efficacy and dosage of drugs. Moreover, model simulations in combination with computational sensitivity analysis and model optimization have been used to detect new potential drug targets in cancer and metabolic diseases, or to assess the emergence of therapy resistance (Vera et al., 2007; Schoeberl et al., 2009). This strategy can be replicated in lung infection diseases to search for new drug targets or repurpose existing drugs as immunomodulators during lung infection (Wentker et al., 2017), or to optimize the current protocols for antibiotics administration (Schirm et al., 2016). Further, in recent times (Zhou et al., 2017) modeling has been used to assess therapies in a personalized manner (Rosenberg and Restifo, 2015; van de Sant et al., 2017), especially anticancer ones (Gupta et al., 2016). We think there is potential for this in lung infection and

pneumonia, by integrating selected patient unique -omics and physiological parameters into model simulations, and use them to customize treatments.

Mathematical Modeling and Multi-Level Dissection of Bacterial Lung Infection: The Art of Choosing the Right Approach

There is no perfect modeling framework for investigating bacterial (lung) infection in all possible scenarios. This is because the optimality of a modeling strategy will depend on the aim of investigation, the scale and structural complexity of the system to be modeled and the quantity, quality and nature of the experimental data available for its characterization. **Table 2** extends our previously published table (Vera and Wolkenhauer, 2011) and compares the main modeling frameworks here discussed based on a number of important features. We also include some prototypical case studies in bacterial lung infection in which each modeling framework could be most suited. One can see that there is no a modeling approach clearly superior to all the others for every feature analyzed, and therefore the choice of the right model relies often on a tight balance between several of these features (**Table 3**). Moreover, in some cases any of the methodologies described displays the features necessary to model the dynamics of given biological systems, and other modeling

TABLE 2 | Features of different model formalisms analyzed.

| Modeling framework | Realism | Time | Scalability | Computational cost | Complexity | Data usage | Examples |
|--------------------|------------------|------------|-------------|--------------------|------------|------------|---|
| ABM | Phenomenological | Continuous | Large | High | High | Low | Spatial simulation of cell-2-cell interaction and movement in a lung alveolus during infection |
| Boolean | Phenomenological | Discrete | Medium | Low | Low | Medium | Analysis and simulation of the large regulatory network triggered in macrophages after bacteria detection |
| ODE | Mechanistic | Continuous | Small | High | Medium | High | Analysis of fine-tuning of NFkB signaling activation in lung epithelial cells after infection |
| Stochastic | Mechanistic | Continuous | Small | High | High | High | Simulation of dynamics of few bacteria initiating infection in a lung alveolus |
| Fuzzy Logic | Phenomenological | Discrete | Medium | Medium | Low | Medium | Simulation of lung epithelial cell phenotypes with uncertain uncomplete information of activators |

Realism: How close from the real biological mechanism is the representation given by the model; time: Whether the model handle the time as a discrete or continuous variable; scalability: Number of compounds the model can on average handle (small: up to 20 compounds, medium: 20–100, large 100–1,000; computational cost: Time and computational resources demanded for model simulation and analysis; complexity of the models in terms of their structure; data usage: Whether the construction of the model requires low, medium or large amounts of quantitative experimental data for its characterization; examples: Possible applications for each formalism in the context of bacterial lung infection.

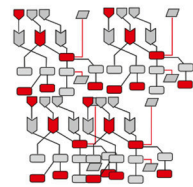
TABLE 3 | Applicability of different model formalisms analyzed into different biological situations.

| Modeling framework | Intracellular circuits | Cell-cell interactions | Host-pathogen interactions | Local tissue interaction | Systemic infection/inflammation |
|--------------------|------------------------|------------------------|----------------------------|--------------------------|---------------------------------|
| ABM | | | | | |
| Boolean | | | | | |
| ODE | | | | | |
| Stochastic | | | | | |
| Fuzzy Logic | | | | | |

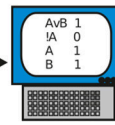
Applicability: This table presents an illustrative guidance to select the best modeling framework to the biological scale of interest. Depending on the scale the applicability of the different frameworks can be poor (black) possible (gray) or appropriate (white).

Informed hybrid models

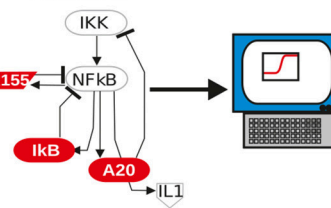
Large regulatory network



Boolean modelling and simulation



Core regulatory network

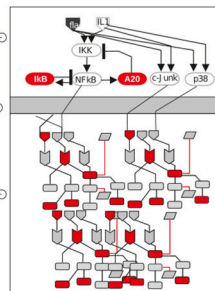


Connected hybrid models

ODE model

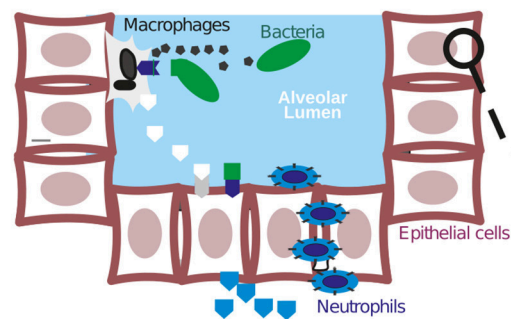
Model-2-model interface

Boolean model



Fully embedded hybrid models

ABM model



ODE model

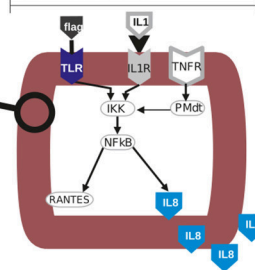


FIGURE 7 | Different strategies for hybrid modeling in bacterial lung infection.

frameworks can be used (See Supplementary Material for further discussion).

In some cases a single modeling approach is not sufficient to deal with some structurally complex systems, and one has to combine different model types into a “hybrid model” (Chiam et al., 2006; Wylie et al., 2006; Wu and Voit, 2009). Agent-based models has become the most used approach in biomedicine for multi-level and multi-scale systems (Chavali et al., 2008). However, other hybrid modeling strategies are implemented by combining modeling approaches with computational and knowledge requirements of different complexities, like Boolean and ODE model together (Figure 7). For example, one can use

the knowledge generated by simulations with a given type of model to parameterize and characterize a second type of model. In this “**informed hybrid models**” there is no formal connection between the models, but one of them is used to design or characterize a second one. For example, In Rex and collaborators simulations on a large Boolean network were used to describe the key regulatory circuits underlying the shift between M1 (classical, LPS-activated, pro inflammatory) M2 (IL4/IL13 activated, anti-inflammatory) macrophage phenotypes (Rex et al., 2016). This information, the key molecular species and their interactions, was used to construct a second ODE model that dissects the fine regulation of this subnetwork.

Another option could be to construct models in different frameworks that are primarily independent, but cross-talk via a few common compounds. An example of this “**connected hybrid models**” could be a combination of an ODE model accounting for a signaling circuit controlling the activation of a number of key transcription factors after bacterial infection (e.g., NF- κ B, p38), connected to a large Boolean network accounting for the activation of dozens to hundreds of transcriptional targets. The connection between both types of models could be done via interface functions accounting for the activation status of the transcription factors (Khan et al., 2014).

Finally in the “**fully embedded hybrid models**” a model in a given formalism is fully integrated in another type of model (Chiam et al., 2006). We think this is an alternative in which ABM could be a suitable option. For example, in multi-scale models accounting for bacterial lung infection one could develop an ABM in which individual bacteria, lung epithelial cells, alveolar macrophages or neutrophils populations are modeled like interacting agents moving within a defined space. The activation, differentiation or apoptotic phenotypes of these agent-cells would be determined by the simulation of embedded Boolean or ODE models, which describe the time-dependent activation of their core intracellular network.

REFERENCES

- Aaby, B., Perumalla, K., and Seal, K. (2010). “Efficient simulation of agent-based models on multi-GPU and multi-core clusters,” in *SIMUTools* (Tennessee).
- Agliari, E., Asti, L., Barra, A., Scrivo, R., Valesini, G., and Wallis, R. S. (2013). Application of a stochastic modeling to assess the evolution of tuberculous and non-tuberculous mycobacterial infection in patients treated with tumor necrosis factor inhibitors. *PLoS ONE* 8:e55017. doi: 10.1371/journal.pone.0055017
- Anderson, C. S., DeDiego, M. L., Topham, D. J., and Thakar, J. (2016). Boolean modeling of cellular and molecular pathways involved in influenza infection. *Comput. Math. Methods. Med.* 2016:7686081. doi: 10.1155/2016/7686081
- Apiwatanakul, N., Thomas, P. G., Kuhn, R. E., Herbert, D. R., and McCullers, J. A. (2014). Helminth infections predispose mice to pneumococcal pneumonia but not to other pneumonic pathogens. *Med. Microbiol. Immunol.* 203, 357–364. doi: 10.1007/s00430-014-0344-3
- Ashall, L., Horton, C. A., Nelson, D. E., Paszek, P., Harper, C. V., Sillitoe, K., et al. (2009). Pulsatile stimulation determines timing and specificity of NF- κ B-dependent transcription. *Science* 324, 242–246. doi: 10.1126/science.1164860
- Auffray, C., Chen, Z., and Hood, L. (2009). Systems medicine: the future of medical genomics and healthcare. *Genome Med.* 1:2. doi: 10.1186/gm2
- Bakstad, D., Adamson, A., Spiller, D. G., and White, M. R. H. (2012). Quantitative measurement of single cell dynamics. *Curr. Opin. Biotechnol.* 23, 103–109. doi: 10.1016/j.copbio.2011.11.007
- Ballesta, A., Dulong, S., Abbara, C., Cohen, B., Okyar, A., Clairambault, J., et al. (2011). A combined experimental and mathematical approach for molecular-based optimization of irinotecan circadian delivery. *PLoS Comput. Biol.* 7:e1002143. doi: 10.1371/journal.pcbi.1002143
- Bendall, S. C., and Nolan, G. P. (2012). From single cells to deep phenotypes in cancer. *Nat. Biotechnol.* 30, 639–647. doi: 10.1038/nbt.2283
- Bertaux, F., Stoma, S., Drasdo, D., and Batt, G. (2014). Modeling dynamics of cell-to-cell variability in TRAIL-induced apoptosis explains fractional killing and predicts reversible resistance. *PLoS Comput. Biol.* 10:e1003893. doi: 10.1371/journal.pcbi.1003893
- Bianchi, C., Cirillo, P., Gallegati, M., and Vagliasindi, P. A. (2007). Validating and calibrating agent-based models: a case study. *Comput. Econ.* 30, 245–264. doi: 10.1007/s10614-007-9097-z
- Bogle, G., and Dunbar, P. R. (2010). Agent-based simulation of T-cell activation and proliferation within a lymph node. *Immunol. Cell Biol.* 88, 172–179. doi: 10.1038/icb.2009.78
- Bressloff, P. C. (2017). *Stochastic Processes in Cell Biology*. Springer. Available online at: <http://www.springer.com/us/book/9783319084879>
- Brown, B. N., Price, I. M., Toapanta, F. R., DeAlmeida, D. R., Wiley, C. A., Ross, T. M., et al. (2011). An agent-based model of inflammation and fibrosis following particulate exposure in the lung. *Math. Biosci.* 231, 186–196. doi: 10.1016/j.mbs.2011.03.005
- Brown, R. A., Spina, D., and Page, C. P. (2008). Adenosine receptors and asthma. *Br. J. Pharmacol.* 153 (Suppl. 1), S446–S456. doi: 10.1038/bjp.2008.22
- Castillo-Montiel, E., Chimal-Eguía, J. C., Tello, J. I., Piñon-Zarate, G., Herrera-Enríquez, M., and Castell-Rodríguez, A. E. (2015). Enhancing dendritic cell immunotherapy for melanoma using a simple mathematical model. *Theor. Biol. Med. Model.* 12:11. doi: 10.1186/s12976-015-0007-0
- Celli, S., Day, M., Müller, A. J., Molina-Paris, C., Lythe, G., and Bouso, P. (2012). How many dendritic cells are required to initiate a T-cell response? *Blood* 120, 3945–3948. doi: 10.1182/blood-2012-01-408260
- Chavali, A. K., Gianchandani, E. P., Tung, K. S., Lawrence, M. B., Peirce, S. M., and Papin, J. A. (2008). Characterizing emergent properties of immunological systems with multi-cellular rule-based computational modeling. *Trends Immunol.* 29, 589–599. doi: 10.1016/j.it.2008.08.006
- Cheepsattayakorn, A., and Cheepsattayakorn, R. (2014). Parasitic pneumonia and lung involvement. *BioMed. Res. Int.* 2014:874021. doi: 10.1155/2014/874021
- Chen, K., and Kolls, J. K. (2013). T Cell-mediated host immune defenses in the lung. *Annu. Rev. Immunol.* 31, 605–633. doi: 10.1146/annurev-immunol-032712-100019
- Chernyavsky, I. L., Croisier, H., Chapman, L. A. C., Kimpton, L. S., Hiorns, J. E., Brook, B. S., et al. (2014). The role of inflammation resolution speed in airway smooth muscle mass accumulation in asthma: insight from a theoretical model. *PLoS ONE* 9:e90162. doi: 10.1371/journal.pone.0090162
- Chiam, K.-H., Tan, C. M., Bhargava, V., and Rajagopal, G. (2006). Hybrid simulations of stochastic reaction-diffusion processes for modeling intracellular signaling pathways. *Phys. Rev. E Stat. Nonlin. Soft Matt. Phys.* 74(5 Pt 1):051910. doi: 10.1103/PhysRevE.74.051910
- Choi, H. J., Seo, C. H., Park, S. H., Yang, H., Do, K. H., Kim, J., et al. (2011). Involvement of Epidermal Growth Factor Receptor-Linked Signaling Responses in *Pseudomonas fluorescens*-Infected Alveolar Epithelial Cells V. *Infect. Immun.* 79, 1998–2005. doi: 10.1128/IAI.01232-10

AUTHOR CONTRIBUTIONS

All authors listed have made a substantial, direct and intellectual contribution to the work, and approved it for publication. MC and PW wrote the section of the boolean section, GS wrote the section of ODE and ABM, JV wrote the introduction and stochastic models section. XL provided a critical review on the whole manuscript. All authors wrote the discussion and made the final correction to the article.

ACKNOWLEDGMENTS

This work was funded by the German Federal Ministry of Education and Research [BMBF; projects e:Bio-miRSys (0316175A) and e:Med-CAPSyS (01ZX1304F)].

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: <http://journal.frontiersin.org/article/10.3389/fphys.2017.00645/full#supplementary-material>

- Conces, D. J. (1998). Bacterial pneumonia in immunocompromised patients. *J. Thorac. Imaging* 13, 261–270. doi: 10.1097/00005382-199810000-00006
- Contoli, M., Bousquet, J., Fabbri, L. M., Magnussen, H., Rabe, K. F., Siafakas, N. M., et al. (2010). The small airways and distal lung compartment in asthma and COPD: a time for reappraisal. *Allergy* 65, 141–151. doi: 10.1111/j.1398-9995.2009.02242.x
- Coward, W. R., Okayama, Y., Sagara, H., Wilson, S. J., Holgate, S. T., and Church, M. K. (2002). NF-kappa B and TNF-alpha: a positive autocrine loop in human lung mast cells? *J. Immunol. Baltim. Md.* 169, 5287–5293. doi: 10.4049/jimmunol.169.9.5287
- Crépieux, P., Kwon, H., Leclerc, N., Spencer, W., Richard, S., Lin, R., et al. (1997). I kappaB alpha physically interacts with a cytoskeleton-associated protein through its signal response domain. *Mol. Cell Biol.* 17, 7375–7385. doi: 10.1128/MCB.17.12.7375
- Csóka, B., Németh, Z. H., Rosenberger, P., Eltzschig, H. K., Spolarics, Z., Pacher, P., et al. (2010). A2B adenosine receptors protect against sepsis-induced mortality by dampening excessive inflammation. *J. Immunol. Baltim.* 185, 542–550. doi: 10.4049/jimmunol.0901295
- D'Agata, E. M. C., Magal, P., Olivier, D., Ruan, S., and Webb, G. F. (2007). Modeling antibiotic resistance in hospitals: the impact of minimizing treatment duration. *J. Theor. Biol.* 249, 487–499. doi: 10.1016/j.jtbi.2007.08.011
- Dix, A., Vlaic, S., Guthke, R., and Linde, J. (2016). Use of systems biology to decipher host-pathogen interaction networks and predict biomarkers. *Clin. Microbiol. Infect. Off. Publ. Eur. Soc. Clin. Microbiol. Infect. Dis.* 22, 600–606. doi: 10.1016/j.cmi.2016.04.014
- Dobrzyński, M., Nguyen, L. K., Birtwistle, M. R., von Kriegsheim, A., Blanco Fernández, A., Cheong, A., et al. (2014). Nonlinear signalling networks and cell-to-cell variability transform external signals into broadly distributed or bimodal responses. *J. R. Soc. Interface* 11:20140383. doi: 10.1098/rsif.2014.0383
- Domínguez-Hüttinger, E., Boon, N. J., Clarke, T. B., and Tanaka, R. J. (2017). Mathematical modeling of *Streptococcus pneumoniae* colonization, invasive infection and treatment. *Front. Physiol.* 8:115. doi: 10.3389/fphys.2017.00115
- Duan, W., Lee, K., Herbison, A. E., and Sneyd, J. (2011). A mathematical model of adult GnRH neurons in mouse brain and its bifurcation analysis. *J. Theor. Biol.* 276, 22–34. doi: 10.1016/j.jtbi.2011.01.035
- Dunster, J. L., Byrne, H. M., and King, J. R. (2014). The resolution of inflammation: a mathematical model of neutrophil and macrophage interactions. *Bull. Math. Biol.* 76, 1953–1980. doi: 10.1007/s11538-014-9987-x
- Eberhardt, M., Lai, X., Tomar, N., Gupta, S., Schmeck, B., Steinkasserer, A., et al. (2016). Third-kind encounters in biomedicine: immunology meets mathematics and informatics to become quantitative and predictive. *Methods Mol. Biol.* 1386, 135–179. doi: 10.1007/978-1-4939-3283-2_9
- Elowitz, M. B., Levine, A. J., Siggia, E. D., and Swain, P. S. (2002). Stochastic gene expression in a single cell. *Science* 297, 1183–1186. doi: 10.1126/science.1070919
- Epstein, J. M. (2008). *Why Model?*. Available online at: <http://jasss.soc.surrey.ac.uk/11/4/12.html>
- Eyler, R. F., Mueller, B. A., and Medscape. (2011). Antibiotic dosing in critically ill patients with acute kidney injury. *Nat. Rev. Nephrol.* 7, 226–235. doi: 10.1038/nrneph.2011.12
- Finigan, J. H., Downey, G. P., and Kern, J. A. (2012). Human epidermal growth factor receptor signaling in acute lung injury. *Am. J. Respir. Cell Mol. Biol.* 47, 395–404. doi: 10.1165/rcmb.2012-0100TR
- Gane, J. M., Stockley, R. A., and Sapey, E. (2016). TNF- α Autocrine Feedback Loops in Human Monocytes: The Pro- and Anti-Inflammatory Roles of the TNF- α Receptors Support the Concept of Selective TNFR1 Blockade *in vivo*. *J. Immunol. Res.* 2016:1079851. doi: 10.1155/2016/1079851
- Gillard, J. J., Laws, T. R., Lythe, G., and Molina-París, C. (2014). Modeling early events in *Francisella tularensis* pathogenesis. *Front. Cell Infect. Microbiol.* 4:169. doi: 10.3389/fcimb.2014.00169
- Gillespie, D. T. (1992). A rigorous derivation of the chemical master equation. *Phys. Stat. Mech. Its Appl.* 188, 404–425. doi: 10.1016/0378-4371(92)90283-V
- Grundmann, H., and Hellriegel, B. (2006). Mathematical modelling: a tool for hospital infection control. *Lancet Infect. Dis.* 6, 39–45. doi: 10.1016/S1473-3099(05)70325-X
- Guebel, D. V., Schmitz, U., Wolkenhauer, O., and Vera, J. (2012). Analysis of cell adhesion during early stages of colon cancer based on an extended multi-valued logic approach. *Mol. Biosyst.* 8, 1230–1242. doi: 10.1039/c2mb05277f
- Guo, B., Abdelraouf, K., Ledesma, K. R., Chang, K.-T., Nikolaou, M., and Tam, V. H. (2011). Quantitative impact of neutrophils on bacterial clearance in a murine pneumonia model. *Antimicrob. Agents Chemother.* 55, 4601–4605. doi: 10.1128/AAC.00508-11
- Gupta, S. K., Jaitly, T., Schmitz, U., Schuler, G., Wolkenhauer, O., and Vera, J. (2016). Personalized cancer immunotherapy using Systems Medicine approaches. *Brief Bioinform.* 17, 453–467. doi: 10.1093/bib/bbv046
- Gustafsson, L., and Sternad, M. (2013). When can a deterministic model of a population system reveal what will happen on average? *Math. Biosci.* 243, 28–45. doi: 10.1016/j.mbs.2013.01.006
- Haack, F., Burrage, K., Redmer, R., and Uhrmacher, A. M. (2013). Studying the role of lipid rafts on protein receptor bindings with cellular automata. *IEEE/ACM Trans. Comput. Biol. Bioinform.* 10, 760–770. doi: 10.1109/TCBB.2013.40
- Hanson, S. M., Cleghorn, W. M., Francis, D. J., Vishnivetskiy, S. A., Raman, D., Song, X., et al. (2007). Arrestin mobilizes signaling proteins to the cytoskeleton and redirects their activity. *J. Mol. Biol.* 368, 375–387. doi: 10.1016/j.jmb.2007.02.053
- Hoffmann, A., Levchenko, A., Scott, M. L., and Baltimore, D. (2002). The IkappaB-NF-kappaB signaling module: temporal control and selective gene activation. *Science* 298, 1241–1245. doi: 10.1126/science.1071914
- Ibargüen-Mondragón, E., Mosquera, S., Cerón, M., Burbano-Rosero, E. M., Hidalgo-Bonilla, S. P., Esteva, L., et al. (2014). Mathematical modeling on bacterial resistance to multiple antibiotics caused by spontaneous mutations. *Biosystems* 117, 60–67. doi: 10.1016/j.biosystems.2014.01.005
- Jack, J., Wambaugh, J. F., and Shah, I. (2011). Simulating quantitative cellular responses using asynchronous threshold Boolean network ensembles. *BMC Syst. Biol.* 5:109. doi: 10.1186/1752-0509-5-109
- Joseph, C., Togawa, Y., and Shindo, N. (2013). Bacterial and viral infections associated with influenza. *Influenza Other Respir. Viruses.* 7(Suppl. 2), 105–113. doi: 10.1111/irv.12089
- Kaern, M., Elston, T. C., Blake, W. J., and Collins, J. J. (2005). Stochasticity in gene expression: from theories to phenotypes. *Nat. Rev. Genet.* 6, 451–464. doi: 10.1038/nrg1615
- Kang, C. C., Chuang, Y. J., Tung, K. C., Chao, C. C., Tang, C. Y., Peng, S. C., et al. (2011). A genetic algorithm-based Boolean delay model of intracellular signal transduction in inflammation. *BMC Bioinformatics* 12(Suppl. 1):S17. doi: 10.1186/1471-2105-12-S1-S17
- Kang, S., Kahan, S., McDermott, J., Flann, N., and Shmulevich, I. (2014). Biocellion: accelerating computer simulation of multicellular biological system models. *Bioinforma. Oxf. Engl.* 30, 3101–3108. doi: 10.1093/bioinformatics/btu498
- Kauffman, S. A. (1969). Metabolic stability and epigenesis in randomly constructed genetic nets. *J. Theor. Biol.* 22, 437–467. doi: 10.1016/0022-5193(69)90015-0
- Kauffman, S. A. (1993). *The Origins of Order: Self-organization and Selection in Evolution*. Oxford University Press.
- Khan, F. M., Schmitz, U., Nikolov, S., Engelmann, D., Pützer, B. M., Wolkenhauer, O., et al. (2014). Hybrid modeling of the crosstalk between signaling and transcriptional networks using ordinary differential equations and multi-valued logic. *Biochim. Biophys. Acta* 1844(1 Pt B), 289–298. doi: 10.1016/j.bbapap.2013.05.007
- Klipp, E., Liebermeister, W., Wierling, C., Kowald, A., Lehrach, H., and Herwig, R. (2009). *Systems Biology: A Textbook*. John Wiley & Sons.
- Krishna, S., Jensen, M. H., and Sneppen, K. (2006). Minimal model of spiky oscillations in NF-kappaB signaling. *Proc. Natl. Acad. Sci. U.S.A.* 103, 10840–10845. doi: 10.1073/pnas.0604085103
- Layek, R., Datta, A., Bittner, M., and Dougherty, E. R. (2011). Cancer therapy design based on pathway logic. *Bioinforma. Oxf. Engl.* 27, 548–555. doi: 10.1093/bioinformatics/btq703
- Levine, J. H., Lin, Y., and Elowitz, M. B. (2013). Functional roles of pulsing in genetic circuits. *Science* 342, 1193–1200. doi: 10.1126/science.1239999
- Levy, P. S., and Green, G. M. (1968). A stochastic model of the bactericidal activity of the lung. *J. Theor. Biol.* 21, 103–112. doi: 10.1016/0022-5193(68)90063-5
- Li, N., Ren, A., Wang, X., Fan, X., Zhao, Y., Gao, G. F., et al. (2015). Influenza viral neuraminidase primes bacterial coinfection through TGF- β -mediated expression of host cell receptors. *Proc. Natl. Acad. Sci. U.S.A.* 112, 238–243. doi: 10.1073/pnas.1414422112

- Lidke, D. S., and Wilson, B. S. (2009). Caught in the act: quantifying protein behaviour in living cells. *Trends Cell Biol.* 19, 566–574. doi: 10.1016/j.tcb.2009.08.004
- Lo, W.-C., Arsenescu, R. I., and Friedman, A. (2013). Mathematical model of the roles of T cells in inflammatory bowel disease. *Bull. Math. Biol.* 75, 1417–1433. doi: 10.1007/s11538-013-9853-2
- Lugade, A. A., Bogner, P. N., and Thanavala, Y. (2011). Murine model of chronic respiratory inflammation. *Adv. Exp. Med. Biol.* 780, 125–141. doi: 10.1007/978-1-4419-5632-3_11
- Mathew, S., Bartels, J., Banerjee, I., and Vodovotz, Y. (2014). Global sensitivity analysis of a mathematical model of acute inflammation identifies nonlinear dependence of cumulative tissue damage on host interleukin-6 responses. *J. Theor. Biol.* 358, 132–148. doi: 10.1016/j.jtbi.2014.05.036
- Matzavinos, A., Chaplain, M. A. J., and Kuznetsov, V. A. (2004). Mathematical modelling of the spatio-temporal response of cytotoxic T-lymphocytes to a solid tumour. *Math. Med. Biol. J. IMA* 21, 1–34. doi: 10.1093/imammb/21.1.1
- Mehta, S., Shin, H., Burnett, R., North, T., and Cohen, A. J. (2013). Ambient particulate air pollution and acute lower respiratory infections: a systematic review and implications for estimating the global burden of disease. *Air. Qual. Atmosphere Health* 6, 69–83. doi: 10.1007/s11869-011-0146-3
- Minton, A. P. (2006). How can biochemical reactions within cells differ from those in test tubes? *J. Cell Sci.* 119(Pt 14), 2863–2869. doi: 10.1242/jcs.03063
- Mochan, E., Swigon, D., Ermentrout, G. B., Lukens, S., and Clermont, G. (2014). A mathematical model of intrahost pneumococcal pneumonia infection dynamics in murine strains. *J. Theor. Biol.* 353, 44–54. doi: 10.1016/j.jtbi.2014.02.021
- Mogensen, T. H. (2009). Pathogen recognition and inflammatory signaling in innate immune defenses. *Clin. Microbiol. Rev.* 22, 240–273. doi: 10.1128/CMR.00046-08
- Mukhopadhyay, S., Hoidal, J. R., and Mukherjee, T. K. (2006). Role of TNF α in pulmonary pathophysiology. *Respir. Res.* 7:125. doi: 10.1186/1465-9921-7-125
- Murano, T., Kagawa, Y., and Tsuneda, S. (2014). Mathematical modelling of spatio-temporal cell dynamics in colonic crypts following irradiation. *Cell Prolif.* 47, 347–355. doi: 10.1111/cpr.12110
- Nature Editorial (2016). Back to Earth. *Nature* 530, 253–254. doi: 10.1038/530253b
- Nelson, D. E., Ihekweaba, A. E. C., Elliott, M., Johnson, J. R., Gibney, C. A., Foreman, B. E., et al. (2004). Oscillations in NF- κ B signaling control the dynamics of gene expression. *Science* 306, 704–708. doi: 10.1126/science.1099962
- Neurath, M. F., Finotto, S., and Glimcher, L. H. (2002). The role of Th1/Th2 polarization in mucosal immunity. *Nat. Med.* 8, 567–573. doi: 10.1038/nm0602-567
- Nikolov, S., Lai, X., Liebal, U. W., Wolkenhauer, O., and Vera, J. (2010). Integration of sensitivity and bifurcation analysis to detect critical processes in a model combining signalling and cell population dynamics. *Int. J. Syst. Sci.* 41, 81–105. doi: 10.1080/00207720903147746
- Norman, T. M., Lord, N. D., Paulsson, J., and Losick, R. (2015). Stochastic switching of cell fate in microbes. *Annu. Rev. Microbiol.* 69, 381–403. doi: 10.1146/annurev-micro-091213-112852
- Nouailles, G., Dorhoi, A., Koch, M., Zerrahn, J., Weiner, J., Faé, K. C., et al. (2014). CXCL5-secreting pulmonary epithelial cells drive destructive neutrophilic inflammation in tuberculosis. *J. Clin. Invest.* 124, 1268–1282. doi: 10.1172/JCI72030
- Ohta, A., and Sitkovsky, M. (2001). Role of G-protein-coupled adenosine receptors in downregulation of inflammation and protection from tissue damage. *Nature* 414, 916–920. doi: 10.1038/414916a
- Pahle, J. (2009). Biochemical simulations: stochastic, approximate stochastic and hybrid approaches. *Brief Bioinform.* 10, 53–64. doi: 10.1093/bib/bbn050
- Patel, B. V., Wilson, M. R., O'Dea, K., and Takata, M. (2013). TNF-induced death signaling triggers alveolar epithelial dysfunction in acute lung injury. *J. Immunol. Baltim. Md.* 190, 4274–4282. doi: 10.4049/jimmunol.1202437
- Pękański, J., Zuk, P. J., Kończyk, M., Junkin, M., Kellogg, R., Tay, S., et al. (2013). Spontaneous NF- κ B activation by autocrine TNF α signaling: a computational analysis. *PLoS ONE* 8:e78887. doi: 10.1371/journal.pone.0078887
- Pelaia, G., Vatrella, A., Gallelli, L., Renda, T., Cazzola, M., Maselli, R., et al. (2006). Respiratory infections and asthma. *Respir. Med.* 100, 775–784. doi: 10.1016/j.rmed.2005.08.025
- Pothen, J. J., Poynter, M. E., and Bates, J. H. T. (2015). A computational model of unresolved allergic inflammation in chronic asthma. *Am. J. Physiol. Lung Cell Mol. Physiol.* 308, L384–L390. doi: 10.1152/ajplung.00268.2014
- Rahmandad, H., and Sterman, J. (2008). Heterogeneity and network structure in the dynamics of diffusion: comparing agent-based and differential equation models. *Manag. Sci.* 54, 998–1014. doi: 10.1287/mnsc.1070.0787
- Raue, A., Kreutz, C., Maiwald, T., Bachmann, J., Schilling, M., Klingmüller, U., et al. (2009). Structural and practical identifiability analysis of partially observed dynamical models by exploiting the profile likelihood. *Bioinform. Oxf. Engl.* 25, 1923–1929. doi: 10.1093/bioinformatics/btp358
- Rex, J., Albrecht, U., Ehling, C., Thomas, M., Zanger, U. M., Sawodny, O., et al. (2016). Model-Based Characterization of Inflammatory Gene Expression Patterns of Activated Macrophages. *PLoS Comput. Biol.* 12:e1005018. doi: 10.1371/journal.pcbi.1005018
- Rhodes, D. M., Smith, S. A., Holcombe, M., and Qvarnstrom, E. E. (2015). Computational Modelling of NF- κ B Activation by IL-1RI and Its Co-Receptor TLR2, Predicts a Role for Cytoskeletal Sequestration of I κ B α in Inflammatory Signalling. *PLoS ONE* 10:e0129888. doi: 10.1371/journal.pone.0129888
- Rivas, G., Ferrone, F., and Herzfeld, J. (2004). Life in a crowded world. *EMBO Rep.* 5, 23–27. doi: 10.1038/sj.embor.7400056
- Rosenberg, S. A., and Restifo, N. P. (2015). Adoptive cell transfer as personalized immunotherapy for human cancer. *Science* 348, 62–68. doi: 10.1126/science.aaa4967
- Rüdiger, S. (2014). Stochastic models of intracellular calcium signals. *Phys. Rep.* 534, 39–87. doi: 10.1016/j.physrep.2013.09.002
- Saez-Rodriguez, J., Simeoni, L., Lindquist, J. A., Hemenway, R., Bommhardt, U., Arndt, B., et al. (2007). A logical model provides insights into T cell receptor signaling. *PLoS Comput. Biol.* 3:e163. doi: 10.1371/journal.pcbi.0030163
- Saltelli, A., Tarantola, S., and Campolongo, F. (2000). Sensitivity analysis as an ingredient of modeling. *Stat. Sci.* 15, 377–395. doi: 10.1214/ss/1009213004
- Santos, G., Diaz, M., and Torres, N. V. (2016). Lipid Raft Size and Lipid Mobility in Non-raft Domains Increase during Aging and are Exacerbated in APP/PS1 Mice Model of Alzheimer's Disease. Predictions from an Agent-Based Mathematical Model. *Front. Physiol.* 7:90. doi: 10.3389/fphys.2016.00090
- Savageau, M. A. (1971). Parameter sensitivity as a criterion for evaluating and comparing the performance of biochemical systems. *Nature* 229, 542–544. doi: 10.1038/229542a0
- Savageau, M. A. (2011). Biomedical engineering strategies in system design space. *Ann. Biomed. Eng.* 39, 1278–1295. doi: 10.1007/s10439-010-0220-2
- Schirm, S., Ahnert, P., Wienhold, S., Mueller-Redetzky, H., Nouailles-Kursar, G., Loeffler, M., et al. (2016). A biomathematical model of pneumococcal lung infection and antibiotic treatment in mice. *PLoS ONE* 11:e0156047. doi: 10.1371/journal.pone.0156047
- Schlatter, R., Schmich, K., Avalos Vizcarra, I., Scheurich, P., Sauter, T., Borner, C., et al. (2009). ON/OFF and beyond—a boolean model of apoptosis. *PLoS Comput. Biol.* 5:e1000595. doi: 10.1371/journal.pcbi.1000595
- Schoeberl, B., Pace, E. A., Fitzgerald, J. B., Harms, B. D., Xu, L., Nie, L., et al. (2009). Therapeutically targeting ErbB3: a key node in ligand-induced activation of the ErbB receptor-PI3K axis. *Sci. Signal.* 2:ra31. doi: 10.1126/scisignal.2000352
- Shmulevich, I., and Aitchison, J. D. (2009). Deterministic and stochastic models of genetic regulatory networks. *Methods Enzymol.* 467, 335–356. doi: 10.1016/S0076-6879(09)67013-0
- Shmulevich, I., Dougherty, E. R., Kim, S., and Zhang, W. (2002). Probabilistic Boolean networks: a rule-based uncertainty model for gene regulatory networks. *Bioinformatics* 18, 261–274. doi: 10.1093/bioinformatics/18.2.261
- Silver, L. L. (2011). Challenges of antibacterial discovery. *Clin. Microbiol. Rev.* 24, 71–109. doi: 10.1128/CMR.00030-10
- Smith, A. M., Adler, F. R., Ribeiro, R. M., Gutenkunst, R. N., McAuley, J. L., McCullers, J. A., et al. (2013). Kinetics of coinfection with influenza A virus and *Streptococcus pneumoniae*. *PLoS Pathog.* 9:e1003238. doi: 10.1371/journal.ppat.1003238
- Smith, A. M., McCullers, J. A., and Adler, F. R. (2011). Mathematical model of a three-stage innate immune response to a pneumococcal lung infection. *J. Theor. Biol.* 276, 106–116. doi: 10.1016/j.jtbi.2011.01.052
- Spiller, D. G., Wood, C. D., Rand, D. A., and White, M. R. H. (2010). Measurement of single-cell dynamics. *Nature* 465, 736–745. doi: 10.1038/nature09232
- Starruß, J., de Back, W., Brusch, L., and Deutsch, A. (2014). Morpheus: a user-friendly modeling environment for multiscale and

- multicellular systems biology. *Bioinforma. Oxf. Engl.* 30, 1331–1332. doi: 10.1093/bioinformatics/btt772
- Steinwede, K., Henken, S., Bohling, J., Maus, R., Ueberberg, B., Brumshagen, C., et al. (2012). TNF-related apoptosis-inducing ligand (TRAIL) exerts therapeutic efficacy for the treatment of pneumococcal pneumonia in mice. *J. Exp. Med.* 209, 1937–1952. doi: 10.1084/jem.20120983
- Stern, J. R., Christley, S., Zaborina, O., Alverdy, J. C., and An, G. (2012). Integration of TGF- β and EGFR based signaling pathways using an Agent Based Model of EPITHELIAL RESTITUTION. *Wound Repair Regen.* 20, 862–871. doi: 10.1111/j.1524-475X.2012.00852.x
- Stupka, J. E., Mortensen, E. M., Anzueto, A., and Restrepo, M. I. (2009). Community-acquired pneumonia in elderly patients. *Aging Health* 5, 763–774. doi: 10.2217/ahe.09.74
- Takashima, K., Tateda, K., Matsumoto, T., Iizawa, Y., Nakao, M., and Yamaguchi, K. (1997). Role of tumor necrosis factor alpha in pathogenesis of pneumococcal pneumonia in mice. *Infect. Immun.* 65, 257–260.
- Tay, S., Hughey, J. J., Lee, T. K., Lipniacki, T., Quake, S. R., and Covert, M. W. (2010). Single-cell NF-kappaB dynamics reveal digital activation and analogue information processing. *Nature* 466, 267–271. doi: 10.1038/nature09145
- Terfve, C. D. A., Wilkes, E. H., Casado, P., Cutillas, P. R., and Saez-Rodriguez, J. (2015). Large-scale models of signal propagation in human cells derived from discovery phosphoproteomic data. *Nat. Commun.* 6:8033. doi: 10.1038/ncomms9033
- Thakar, J., Pathak, A. K., Murphy, L., Albert, R., and Cattadori, I. M. (2012). Network model of immune responses reveals key effectors to single and co-infection dynamics by a respiratory bacterium and a gastrointestinal helminth. *PLoS Comput Biol.* 8:e1002345. doi: 10.1371/journal.pcbi.1002345
- Thakar, J., Piliore, M., Kirimanjeswara, G., Harvill, E. T., and Albert, R. (2007). Modeling systems-level regulation of host immune responses. *PLoS Comput Biol.* 3:e109. doi: 10.1371/journal.pcbi.0030109
- Thakar, J., Saadatpour-Moghaddam, A., Harvill, E. T., and Albert, R. (2009). Constraint-based network model of pathogen-immune system interactions. *J. R. Soc Interface* 6, 599–612. doi: 10.1098/rsif.2008.0363
- Tranquillo, R. T., Fisher, E. S., Farrell, B. E., and Lauffenburger, D. A. (1989). A stochastic model for chemosensory cell movement: application to neutrophil and macrophage persistence and orientation. *Math. Comput. Model.* 12:1179. doi: 10.1016/0895-7177(89)90249-5
- Tuchscher, L., Medina, E., Hussain, M., Völker, W., Heitmann, V., Niemann, S., et al. (2011). *Staphylococcus aureus* phenotype switching: an effective bacterial strategy to escape host immune response and establish a chronic infection. *EMBO Mol. Med.* 3, 129–141. doi: 10.1002/emmm.201000115
- Turner, D. A., Paszek, P., Woodcock, D. J., Nelson, D. E., Horton, C. A., Wang, Y., et al. (2010). Physiological levels of TNF α stimulation induce stochastic dynamics of NF-kappaB responses in single living cells. *J. Cell Sci.* 123(Pt 16), 2834–2843. doi: 10.1242/jcs.069641
- Tyson, J. J., Chen, K. C., and Novak, B. (2003). Sniffers, buzzers, toggles and blinkers: dynamics of regulatory and signaling pathways in the cell. *Curr. Opin. Cell Biol.* 15, 221–231. doi: 10.1016/S0955-0674(03)00017-6
- Ullah, M., and Wolkenhauer, O. (2010). Stochastic approaches in systems biology. *Wiley Interdiscip. Rev. Syst. Biol. Med.* 2, 385–397. doi: 10.1002/wsbm.78
- Uwada, J., Yazawa, T., Islam, M. T., Khan, M. R. I., Krug, S. M., Fromm, M., et al. (2017). Activation of muscarinic receptors prevents TNF- α -mediated intestinal epithelial barrier disruption through p38 MAPK. *Cell Signal.* 35, 188–196. doi: 10.1016/j.cellsig.2017.04.007
- van de Sant, A. J. W., de Vries, N. M., Hoozeboom, T. J., and Nijhuis-van der Sanden, M. W. G. (2017). Implementation of a personalized, cost-effective physical therapy approach (Coach2Move) for older adults: barriers and facilitators. *J. Geriatr. Phys. Ther.* doi: 10.1519/JPT.0000000000000140. [Epub ahead of print].
- Van Furth, R. (2012). *Mononuclear Phagocytes: Characteristics, Physiology and Function*. Springer Science & Business Media.
- Vera, D. J., and Wolkenhauer, P. D. O. (2011). “Mathematical tools in cancer signalling systems biology,” in *Cancer Systems Biology, Bioinformatics and Medicine*, eds A. Cesario, and F. Marcus (Springer), 185–212.
- Vera, J., and Wolkenhauer, O. (2008). A system biology approach to understand functional activity of cell communication systems. *Methods Cell Biol.* 90, 399–415. doi: 10.1016/S0091-679X(08)00817-0
- Vera, J., Curto, R., Cascante, M., and Torres, N. V. (2007). Detection of potential enzyme targets by metabolic modelling and optimization: application to a simple enzymopathy. *Bioinforma. Oxf. Engl.* 23, 2281–2289. doi: 10.1093/bioinformatics/btm326
- Vera, J., Torres, N. V., Moles, C. G., and Banga, J. (2003). Integrated nonlinear optimization of bioprocesses via linear programming. *AIChE J.* 49, 3173–3187. doi: 10.1002/aic.690491217
- Vig, M., and Kinet, J.-P. (2009). Calcium signaling in immune cells. *Nat. Immunol.* 10, 21–27. doi: 10.1038/nri.f.220
- Wang, R.-S., Saadatpour, A., and Albert, R. (2012). Boolean modeling in systems biology: an overview of methodology and applications. *Phys. Biol.* 9:055001. doi: 10.1088/1478-3975/9/5/055001
- Webster, B., Eklund, E. H., Agle, L. M., Chyou, S., Ruggieri, R., and Lu, T. T. (2006). Regulation of lymph node vascular growth by dendritic cells. *J. Exp. Med.* 203, 1903–1913. doi: 10.1084/jem.20052272
- Wentker, P., Eberhardt, M., Dreyer, F. S., Bertrams, W., Cantone, M., Griss, K., et al. (2017). An interactive macrophage signal transduction map facilitates comparative analyses of high-throughput data. *J. Immunol. Baltim. Md.* 198, 2191–2201. doi: 10.4049/jimmunol.1502513
- WHO (2017a). *Global Priority List of Antibiotic-Resistant Bacteria to Guide Research, Discovery, and Development of New Antibiotics*. WHO. Available online at: <http://www.who.int/medicines/publications/global-priority-list-antibiotic-resistant-bacteria/en/>
- WHO (2017b). *The Top 10 Causes of Death*. WHO. Available online at: <http://www.who.int/mediacentre/factsheets/fs310/en/>
- Wilkinson, D. J. (2009). Stochastic modelling for quantitative description of heterogeneous biological systems. *Nat. Rev. Genet.* 10, 122–133. doi: 10.1038/nrg2509
- Wolkenhauer, O., Auffray, C., Jaster, R., Steinhoff, G., and Dammann, O. (2013). The road from systems biology to systems medicine. *Pediatr. Res.* 73(4 Pt 2), 502–507. doi: 10.1038/pr.2013.4
- Wood, R. M., Egan, J. R., and Hall, I. M. (2014). A dose and time response Markov model for the in-host dynamics of infection with intracellular bacteria following inhalation: with application to *Francisella tularensis*. *J. R. Soc. Interface* 11:20140119. doi: 10.1098/rsif.2014.0119
- Wu, J., and Voit, E. (2009). Hybrid modeling in biochemical systems theory by means of functional petri nets. *J. Bioinform. Comput. Biol.* 7, 107–134. doi: 10.1142/S0219720009004047
- Wylie, D. C., Hori, Y., Dinner, A. R., and Chakraborty, A. K. (2006). A hybrid deterministic-stochastic algorithm for modeling cell signaling dynamics in spatially inhomogeneous environments and under the influence of external fields. *J. Phys. Chem. B.* 110, 12749–12765. doi: 10.1021/jp056231f
- Yuri, A. K. (2017). *Elements of Applied Bifurcation Theory*. Springer. Available online at: <http://www.springer.com/gb/book/9781475724219>
- Zhang, P., Chen, H., Liu, X., and Zhang, Z. (2015). An iterative multi-objective particle swarm optimization-based control vector parameterization for state constrained chemical and biochemical engineering problems. *Biochem. Eng. J.* 103, 138–151. doi: 10.1016/j.bej.2015.07.004
- Zhou, Y., Wang, Y., Wu, Y., Huang, C., Yan, H., Zhu, W., et al. (2017). Individualized dual antiplatelet therapy based on platelet function testing in patients undergoing percutaneous coronary intervention: a meta-analysis of randomized controlled trials. *BMC Cardiovasc. Disord.* 17:6. doi: 10.1186/s12872-017-0582-6
- Zi, Z. (2011). Sensitivity analysis approaches applied to systems biology models. *IET Syst. Biol.* 5:336. doi: 10.1049/iet-syb.2011.0015

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2017 Cantone, Santos, Wentker, Lai and Vera. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

GLOSSARY

Agents: The elements in an ABM that interact in the environment.

Bifurcation analysis: Follow the changes in the qualitative behavior of the simulation of a kinetic model after modifying one parameter.

Bistability: A property that kinetic models can present. It is characterized by two stable steady solutions of the system.

Design space analysis: Identify regions defined by a couple of parameters that show different qualitative behaviors depending on the values of these parameters.

Digital response: Response to stimuli by activating key molecules and phenotypes in an all-or-nothing manner.

Electronic Logic Gates: Elementary electronic block of an ideal circuit that performs logic operations. Usually, each block is characterized by two inputs and one output; logic operations that are performed are: AND, OR, XOR, NAND, XNOR, and NOR. The logic operation with one input and one output line is NOT.

Environment: The open area in which is defined an ABM.

Kinetic model: Mathematical model which consider the time dimension in the simulations.

Mass-Action: Mathematical formalism that represents the velocity of the processes as the product of the elements interacting rose to integer values. These kinetic orders correspond with the stoichiometric values of the interaction.

Model calibration: Searching for the values of the parameters that produce a simulation from a kinetic model that reproduce the dynamic behavior of experimental data.

Model optimization: Searching for the parameter set that best reproduce a specific simulation of interest by a kinetic model.

Parameter identifiability: Problems of certain systems to find a precise value for the parameters given a set of experimental data. This problem ends with a broad uncertainty on the values of some parameters.

Rules: The definition of the possible interaction that the elements can have.

Sensitivity analysis: A mathematical analysis to quantify the effect of the parameters of the model on the response of the simulations.

Symbolic analysis: A qualitative analysis of a kinetic model allowing to identify general patterns without giving specific values to the parameters of the model.

Synchronous and asynchronous algorithms: These methods refer to the update of the nodes' states. In the first case, all nodes are updated during each iteration step following the set of functions defined in the model. Asynchronous updated introduce uncertainties typical of biology: during one iteration, only some nodes randomly chosen are updated accordingly to the defined set of functions.

T cells: A branch of the immune response key for the mid-term response to infection.



Toward Multiscale Models of Cyanobacterial Growth: A Modular Approach

Stefanie Westermark and Ralf Steuer*

Fachinstitut für Theoretische Biologie (ITB), Institut für Biologie, Humboldt-Universität zu Berlin, Berlin, Germany

OPEN ACCESS

Edited by:

Alberto Marin-Sanguino,
Technische Universität München,
Germany

Reviewed by:

Monika Heiner,
Brandenburg University of
Technology, Germany
Ganesh Sriram,
University of Maryland College Park,
USA

*Correspondence:

Ralf Steuer
ralf.steuer@hu-berlin.de

Specialty section:

This article was submitted to
Systems Biology,
a section of the journal
Frontiers in Bioengineering and
Biotechnology

Received: 18 February 2016

Accepted: 09 December 2016

Published: 26 December 2016

Citation:

Westermark S and Steuer R (2016)
Toward Multiscale Models of
Cyanobacterial Growth:
A Modular Approach.
Front. Bioeng. Biotechnol. 4:95.
doi: 10.3389/fbioe.2016.00095

Oxygenic photosynthesis dominates global primary productivity ever since its evolution more than three billion years ago. While many aspects of phototrophic growth are well understood, it remains a considerable challenge to elucidate the manifold dependencies and interconnections between the diverse cellular processes that together facilitate the synthesis of new cells. Phototrophic growth involves the coordinated action of several layers of cellular functioning, ranging from the photosynthetic light reactions and the electron transport chain, to carbon-concentrating mechanisms and the assimilation of inorganic carbon. It requires the synthesis of new building blocks by cellular metabolism, protection against excessive light, as well as diurnal regulation by a circadian clock and the orchestration of gene expression and cell division. Computational modeling allows us to quantitatively describe these cellular functions and processes relevant for phototrophic growth. As yet, however, computational models are mostly confined to the inner workings of individual cellular processes, rather than describing the manifold interactions between them in the context of a living cell. Using cyanobacteria as model organisms, this contribution seeks to summarize existing computational models that are relevant to describe phototrophic growth and seeks to outline their interactions and dependencies. Our ultimate aim is to understand cellular functioning and growth as the outcome of a coordinated operation of diverse yet interconnected cellular processes.

Keywords: photosynthesis, cyanobacteria, whole-cell models, flux balance analysis (FBA), circadian clock, CO₂-concentrating mechanisms (CCMs), network reconstruction, metabolism

1. INTRODUCTION

Almost all life on our planet ultimately depends on harvesting the light energy provided by the sun and the subsequent conversion of atmospheric CO₂ and other inorganic nutrients into the building blocks of life. As one of the key inventions in evolution, oxygenic photosynthesis has transformed life on Earth and dominates the Earth's primary productivity today (Lane, 2002; Morton, 2009). Beyond their evolutionary and ecological importance, phototrophic organisms are an essential resource for humankind and provide almost all food, feed, and fiber required to sustain human life on this planet with more than 7 billion inhabitants. Many of our strategies to master the challenges of the 21st century will inevitably rely on the growth of phototrophic organisms. Making better use of the sun's light energy while avoiding past mistakes of industrial agriculture related to water usage, energy expenditure, eutrophication, and land use are necessary steps for a sustainable future.

Phototrophic microorganisms, in particular cyanobacteria, hold great promise as a renewable resource. Cyanobacteria are able to grow with high yield under adverse conditions and their cultivation does not rely on traditional farmland or fresh water. To make use of the biotechnological potential of cyanobacteria, however, requires further understanding of the organization of phototrophic growth. While many aspects of phototrophic growth are well known and many details of photosynthetic functioning have been unraveled by decades of active research, it still remains a considerable challenge to understand the individual cellular processes in the context of a living cell.

To this end, the construction of computational models of cellular processes offers the possibility to investigate the emergent properties that arise from interacting processes. Corresponding to the path of experimental research, however, to date almost all computational models involving cyanobacterial functioning and growth focus on the inner workings of individual processes, such as the path of electrons in photosystem II or the functioning of the circadian clock. But phototrophic growth is an organismic property. It is not so much an individual process that gives rise to cellular growth, rather it is the interplay of individual processes that bears reproduction and growth of living cells.

In this contribution, we seek to provide an overview on processes relevant to cyanobacterial growth and summarize the available computational models thereof. Our aim is not encyclopedic, that is, we do not aim for a comprehensive account of all available models. Rather, we seek to focus on representative models that may contribute to our understanding of the functioning of a cell as a whole. Our focus are also not the, albeit important, minutiae of individual processes and models, but rather how they can be collated into a coherent whole. Our starting point is a set of existing computational descriptions of cellular processes and their possible interactions. Our ultimate goal is to describe cellular adaptation, cellular resource allocation, and phototrophic growth in complex environments. Or, as more eloquently put by Neidhardt (1999) already more than 15 years ago: “We must solve the cell. That is, we must do our best to design a computer-based model that can predict overall cell behavior for steady states of growth and for transitions between steady states. The model will at first be crude, inaccurate, and a complete failure at some tasks. With increasing refinement based on additional experimental data, the model should gradually improve. Importantly, the model will guide experimental inquiry by indicating areas of inadequate, insufficient, or incorrect information. Vitality, it is only through such modeling of whole-system behavior—that is, of growth—that one will learn how near and how far our knowledge takes us toward understanding the living cell.”

Our premise is that sophisticated computational models are already available for many of the processes that underlie phototrophic growth. Modeling their interactions, however, is still no trivial task. First, most models focus on the inner workings of the processes they describe—and therefore often do not describe key variables that govern the interaction with other processes. Second, the various subprocesses and time scales involved in a computational description of phototrophic growth typically require the use of different mathematical and computational concepts, which cannot always be easily reconciled within a

single computational description. We seek to summarize these different computational descriptions and aim to highlight common variables and interactions. Importantly, we do not necessarily aim at a single unified model that encompassed all aspects of a growing cell. Rather, we argue for a modular approach—a growing set of models that describe aspects of cyanobacterial growth on different temporal and spatial scales. Depending on the research question, and the temporal and spatial scales involved in this particular research question, different descriptions of cyanobacterial functioning may be chosen—and utilized to derive the emergent properties of cellular growth by putting the parts together.

2. MODELING PHOTOTROPHIC GROWTH: AN OVERVIEW

Cellular growth is an organismic process that arises from a coordinated interplay of cellular functions. In the following, we briefly describe the key processes relevant to cyanobacterial functioning and growth in complex environments. An overview is provided in **Figure 1**.

Survival and growth of (most) cyanobacteria begins with the absorption of photons facilitated by large light-harvesting antennae, the phycobilisomes, and chlorophyll *a*. The energy harvested from sunlight drives water splitting at photosystem II (PSII). Electrons, derived from water, are provided to the electron transport chain (ETC) and molecular oxygen is released as the byproduct of photosynthesis.

The ETC consists of a number of large protein complexes, mostly located in the thylakoid membrane. Electrons are transferred along the ETC, ultimately resulting in the regeneration of adenosine triphosphate (ATP) and reduced nicotinamide adenine dinucleotide phosphate (NADPH) as energy carrier and reductant, respectively. The functioning of the photosystems and the ETC are complex biophysical processes and objects of intense research. The respective processes are characterized by fast time scales and transitions between a large number of possible states. While a number of detailed computational models of these processes are available, often with a focus on photosystem II, the respective models typically do not describe regeneration of ATP and NADPH, and hence are not straightforwardly connected to other cellular functions.

The ATP and NADPH regenerated by the photosynthetic light reactions play a crucial role for almost all other cellular processes. Beyond their role as energy donor and reductant, they also serve as important signaling compounds to convey information about the intracellular state. Regenerated ATP and NADPH are utilized to assimilate atmospheric carbon dioxide (CO₂). Cyanobacteria possess mechanisms to concentrate inorganic carbon in the vicinity of the CO₂-fixing enzyme, the ribulose-1,5-bisphosphate carboxylase/oxygenase (RuBisCO), making use of bacterial microcompartments known as carboxysomes. Compared to the light reactions, the relevant time scales of the so-called dark reactions are significantly slower. Modeling of CO₂-concentrating mechanisms (CCMs) typically involves consideration of diffusion and spatial structure.

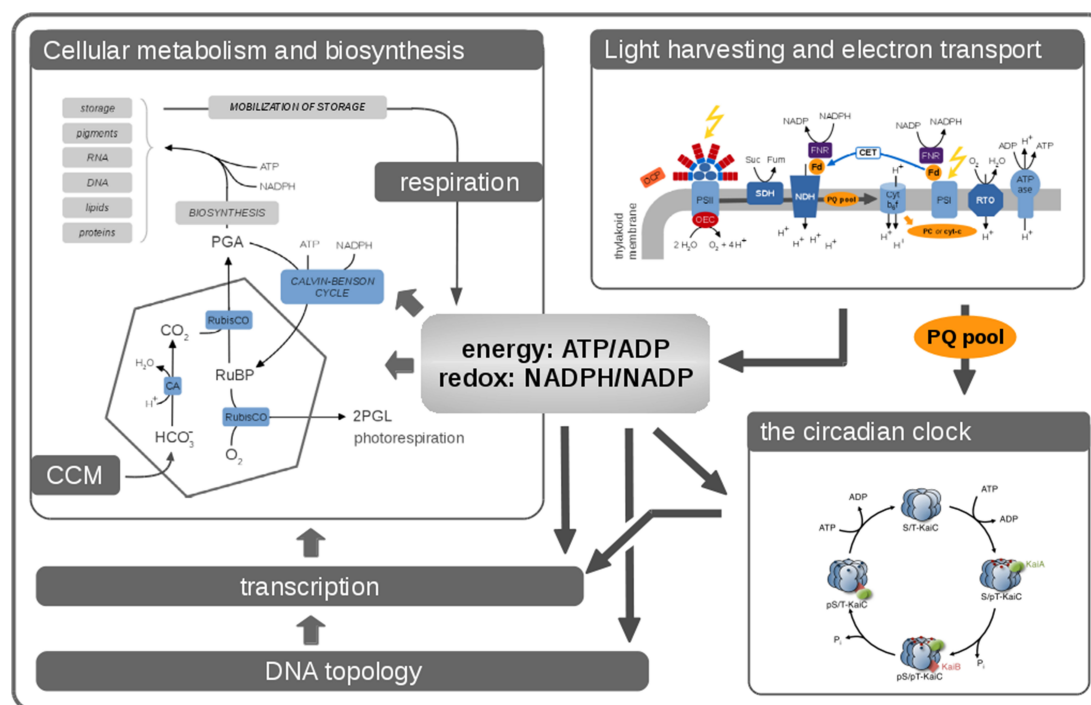


FIGURE 1 | Phototrophic growth is an emergent property that arises from interacting cellular processes. Cellular growth can be described by considering these processes as interacting modules. Many aspects related to cellular growth are reasonably well understood, and detailed computational models already exist. Nonetheless, it remains a considerable challenge to integrate these diverse models into a coherent whole. In this contribution, we seek to describe the processes that are relevant for cyanobacterial functioning and growth. Key aspects are the photosynthetic light reactions, providing energy and reductants for cellular metabolism, cellular respiration in the absence of light, CO₂-concentrating mechanisms (CCMs), transcription, and global energy-induced changes in DNA topology, as well as the circadian clock and its integration into diverse layers of cellular regulation.

The carbon assimilated by the enzyme RuBisCO serves as a substrate to synthesize new cell components, including storage compounds and substrates for cellular respiration. Cellular metabolism involves several hundreds of biochemical reactions, catalyzed by enzymes, as well as spontaneous interconversions, transport, and diffusion processes. From a computational perspective, a description of cellular metabolism must involve several spatial and temporal scales. Detailed models for the action of individual enzymes, such as RuBisCO, exist, based on detailed enzyme mechanisms and elementary reaction steps. Pathways are typically described by combining, often approximative, kinetics of the involved enzymes into larger kinetic models, described by ordinary differential equations (ODEs).

While current kinetic models of metabolism rarely involve more than a few dozen compounds, cellular metabolism is increasingly analyzed using large-scale metabolic reconstructions and constraint-based computational methods. Metabolic network reconstructions are based upon the predicted gene content deduced from genomic DNA and aim to provide an unbiased and comprehensive account of all interconversions of small molecules inside a single cell or a compartment. Unlike kinetic models, metabolic reconstructions only make use of the stoichiometric properties of the respective interconversions, and manually curated reconstructions have been reported for a number of cyanobacteria (Knoop et al., 2010, 2013; Montagud

et al., 2010; Nogales et al., 2012; Saha et al., 2012; Vu et al., 2012; Yoshikawa et al., 2015). Highly efficient computational methods exist that allow for the analysis of networks that consist of several hundreds of biochemical reactions and other molecular interconversions. These computational methods, however, are challenging to reconcile and integrate into more traditional enzyme kinetic models of metabolism.

In addition to the photosynthetic light reactions and cellular metabolism, cyanobacterial functioning also involves a large number of regulatory processes. Most prominent is the cyanobacterial circadian clock. Unique among all known prokaryotes, cyanobacteria possess a true circadian clock, a self-sustained oscillator that is entrained to an external *zeitgeber*. Since its discovery in the late 1980s, the cyanobacterial circadian clock has been an object of intense research (Pattanayak and Rust, 2014). Early research, however, was mostly focused on the inner workings of the clock, the molecular details of the core clock, and its input pathways. Only recently, interactions between the clock and metabolism and the question how the clock functions within a broader cellular context have been addressed in more detail (Pattanayak and Rust, 2014; Diamond et al., 2015; Shultzaberger et al., 2015).

Correspondingly, a number of quantitative computational models exist that describe the mechanistic details of the cyanobacterial clock, as well as its entrainment to environmental

cues—but as yet only few of these models allow for a straightforward integration into a broader cellular context. While there is increasing evidence how the clock is influenced by, and itself influences, photosynthetic light reactions and metabolism, via sensing metabolic activity (Pattanayak et al., 2015) and redox state (Kim et al., 2012) and controlling transcription regulation, the precise evolutionary role of the clock remains insufficiently understood. Elucidating how the circadian clock interacts with other cellular processes and to integrate models of the cyanobacterial circadian clock into a broader cellular context, with the aim to understand how timing mechanisms affect cellular fitness, is a timely question for further computational research.

Energy metabolism and growth, like all cellular processes, are also dependent on the transcriptional and translational machinery. The transcriptional landscape of commonly cultivated cyanobacteria, such as *Synechocystis* sp. PCC 6803, is increasingly known (Kopf et al., 2014) and a number of studies investigated transcriptional rhythms in the presence of light–dark phases (Lehmann et al., 2013; Beck et al., 2014). Of particular interest is also the role of small regulatory RNA (sRNA) to coordinate cellular processes. For several cyanobacterial strains, most notably *Synechocystis* sp. PCC 6803, substantial sRNA transcription, intragenic transcripts, and antisense transcripts have been reported (Kopf and Hess, 2015). While generic computational models for various possible role of regulatory RNA exist (Legewie et al., 2008), these are currently not integrated within larger computational efforts to understand growth properties and adaptation of cyanobacteria.

As a key driver of cellular functioning and growth, global gene expression is believed to be under direct control of the circadian oscillator, mediated by the topological properties of the cyanobacterial chromosomes. It was shown that the superhelicity of the DNA undergoes rhythmic changes that drive global changes in gene expression (Woelfle et al., 2007; Vijayan et al., 2009). It is further known that the rate of transcription also depends on the local supercoiling status of DNA. Vice versa, supercoiling depends on the cellular energy status, since the extent of supercoiling achieved by the DNA gyrase is strongly dependent on ATP hydrolysis. For heterotrophic organisms, specifically *Escherichia coli*, these observations have led to the proposal of homeostatic control and a feedback loop between the intracellular ATP/ADP ratio, DNA supercoiling, transcription, and again changes in the ATP/ADP ratio (Wijker et al., 1995). Closely related ideas have been put forward in the context of ultradian rhythms in yeast where global partitioning of anabolism and catabolism might be mediated by ATP feedback loop on chromatin architecture (Amariei et al., 2013). In the case of cyanobacteria, a global feedback between cellular energy state, DNA supercoiling, and transcription might mediate between global transcription rhythms, the light reactions as the source of cellular energy, and the circadian clock. Such global feedbacks are currently not explicitly considered in models of cyanobacterial growth and are challenging to implement because of the diverse layers of cellular regulation involved.

Parallel to the efforts of molecular biology to understand the mechanistic and biophysical basis of the processes involved in

phototrophic growth, there is a rich history of phenomenological phytoplanktonic growth models. Phenomenological growth formulations are typically employed in models of marine ecosystems and food webs, as well as biogeochemical models to understand the global response of ecosystems to environmental changes. Phenomenological growth models often employ Monod-type equations to describe uptake of a limiting nutrient. Following the early work of Droop (1968), also more sophisticated approaches exist to describe variable internal quotas, see Droop (1983) for an overview. The dynamics of simple phytoplankton growth models are typically based on empirical parameter fitting, rather than an outcome of the underlying cell physiology, and involve strong simplifications, such as using a constant carbon-to-nitrogen (C:N) stoichiometry and absence of photoacclimation (Ayata et al., 2013). It has been pointed out that a major shortcoming of such models is their limited ability to produce true emergence in marine ecosystem models (Allen and Polimene, 2011). Specifically, these models do not evolve to new states not already incorporated in their formulation that makes them unsuitable to properly predict ecosystem changes under changing environmental conditions. As argued by Allen and Polimene (2011), the path forward is to place more emphasis on the underlying intracellular processes—resulting in physiological growth formulations that allow for trade-offs between resource allocations of physiological activities, and hence the possibility to produce biogeochemical and ecological dynamics as emergent properties. Preliminary models, albeit still limited, that combine a detailed description of photosynthesis and phytoplankton growth are already available (Kroon and Thoms, 2006).

In the following, we seek to discuss selected computational models related to cyanobacterial functioning and growth in more detail. Our view is that cyanobacterial physiology depends on interacting cellular processes that can be interpreted as functional “modules”, such as the photosynthetic light reactions and the ETC, carbon uptake mechanisms, cellular metabolism, the circadian clock, as well as the transcriptional and translational machinery and its regulation. For many of these modules, reasonable computational descriptions already exist, whereas other processes, for example, the coordination of cell cycle events in relation to metabolism (Asato, 2005, 2006), have not yet been subject to computational studies.

Our aim is to highlight the common variables and known interactions between the processes relevant to cyanobacterial functioning and growth. In this respect, a particular challenge is the wide range of computational approaches and methods used. Models of cellular processes may take many forms, ranging from spatial versus non-spatial, stochastic versus deterministic, population level versus single cell level, and continuous versus discrete descriptions. See **Figure 2** for an overview. Notwithstanding the technical challenges, we believe that the integration of different aspects of cellular growth, and their respective computational representation, is a prerequisite toward understanding the living cell. We seek to understand how phototrophic growth functions and how it is regulated. How does the coordination of physiological functions work in order to synthesize the right macromolecules at the right time? Which

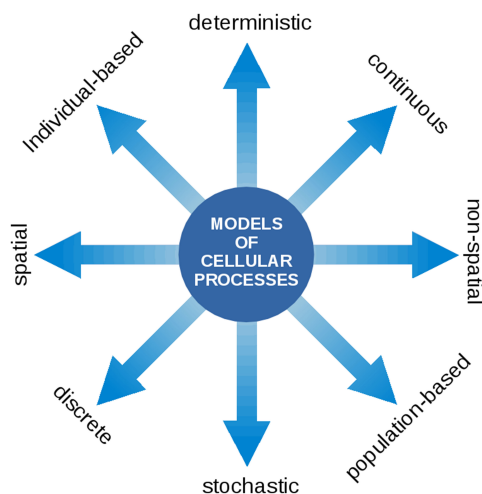


FIGURE 2 | Models of cellular processes are highly diverse and may involve a wide range of computational concepts and methodologies.

At the core of the modeling process is a translation of a biological processes into a formal (mathematical) language. Once this translation is established, the model can be interrogated using the tools of mathematical and computational analysis. The most prevalent representations of cellular processes described in this contribution make use of deterministic ordinary differential equations (ODEs) to describe the time-dependent dynamics of continuous intracellular concentrations, typically on the population level. In the following, such models are denoted as kinetic models and may either make use of heuristic approximate rate equations or rate equations derived from explicit biochemical mechanisms. Models of CCMs typically involve a spatial component and originate from a description based on partial differential equations (PDEs). Models of the light reactions frequently describe transitions between discrete states that occur with a certain state-dependent probability. Flux balance models consists of a set of linear relationships (linear inequality constraints) between variables and make use of linear programming (LP), a method to identify the optimum of a linear objective function. For a more detailed overview on model types, see also Steuer and Junker (2009).

level of detail is required to describe cellular growth? What are the variables and time scales involved?

3. MODELS OF THE PHOTOSYNTHETIC LIGHT REACTIONS

Phototrophic growth begins with the absorption of light and its conversion into chemical energy. Despite a number of open questions and the need for further research, many of the fundamental properties of oxygenic photosynthesis have been elucidated in the past century. Owing to the fact that cyanobacteria are the evolutionary ancestors of modern-day chloroplasts, the organization of their photosynthetic ETC is essentially identical to that in algae and green plants (Vermaas, 2001).

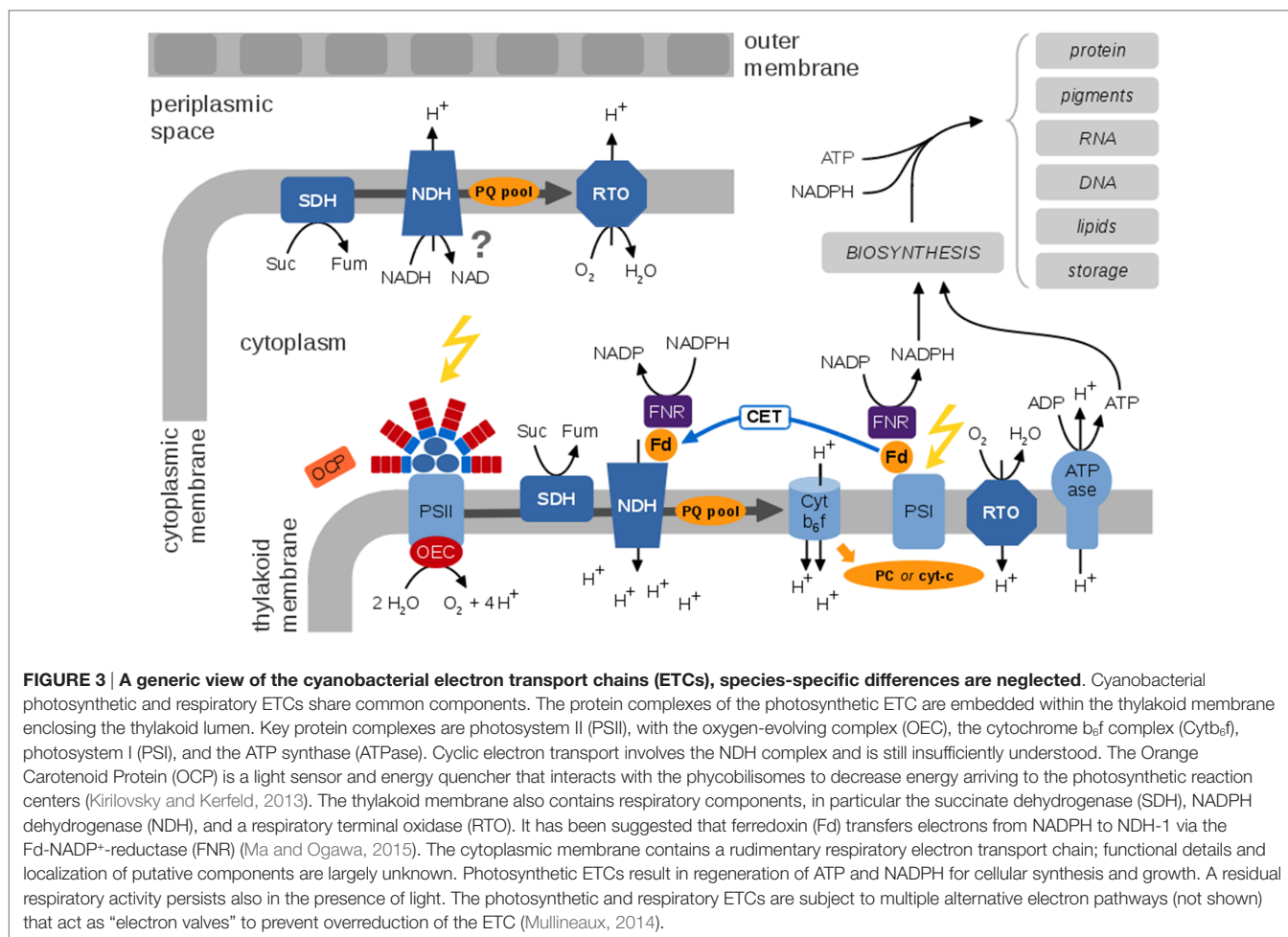
In most cyanobacteria, light harvesting is facilitated by large antenna complexes, the phycobilisomes. Phycobilisomes are attached to the cytoplasmic surface of the thylakoid membrane (Mullineaux, 2014). The detailed composition of phycobilisomes is strain specific and depends on light quality, denoted as complementary chromatic adaptation. The energy absorbed

by the phycobilisomes is transferred to either photosystem II or photosystem I, or dissipated as heat or fluorescence. The protein complexes of the photosynthetic ETC are embedded within the thylakoid membrane. The key proteins complexes responsible for photosynthetic electron transport are Photosystem II (PSII), the Cytochrome b_6f complex (Cyt b_6f), Photosystem I (PSI), and ATP synthase (ATPase). See **Figure 3** for an overview.

PSII splits water and reduces the plastoquinone (PQ) pool. The latter mediates the transport of electrons from PSII to Cyt b_6f . At Cyt b_6f , electrons are transferred to a soluble electron carrier on the luminal side of the thylakoid membrane, either plastocyanine (PC) or cytochrome- c (cyt- c). At PSI, electrons are transferred to ferredoxin and eventually to NADPH using light-induced excitation of the PSI reaction center (linear electron transport, LET). Alternatively, electrons from the excited PSI state can be transferred back to PQ and Cyt b_6f (cyclic electron transport, CET), details of CET are still under debate and insufficiently understood. Photosynthetic electron flow results in a protein gradient across the thylakoid membrane that drives regeneration of ATP by the ATPase.

Unique to cyanobacteria, as opposed to plants and microalgae, is the combination of oxygenic photosynthesis and respiration in the same membrane system using intersecting ETCs and common components (Vermaas, 2001). The respiratory ETC involves the succinate dehydrogenase (SDH), the NADPH dehydrogenase (NDH-1), and terminal oxidases. The PQ pool, the Cyt b_6f complex, and PC (or cyt- c) as soluble electron carrier are involved in respiratory as well as photosynthetic electron transport. While photosynthesis exclusively takes place in the thylakoid membrane, a rudimentary respiratory chain is also present in the plasma membrane (Schultze et al., 2009).

From a computational perspective, photosynthesis in cyanobacteria and microalgae can be described on different levels of complexity. Basic models are closely related to overall growth models in ecology—and typically to reproduce the production of oxygen and the photosynthesis–irradiance (PI) curve of cyanobacteria and microalgae. Early models were derived by Crill (1977), Megard et al. (1984), Eilers and Peeters (1988), and Zonneveld (1998) among others. These models make use of a highly simplified photosynthetic factory or photosynthetic unit (PSU) that encompasses PSII, PSI, and the ETCs. See **Figure 4** for an example. The resulting differential equations for the dependency of photosynthesis on light intensity can often be solved analytically, with a solution analogous to the Haldane equation—an enzyme kinetic equation that was derived for substrates with inhibitory effects at high concentrations. Simple three-state models are suitable to describe basic features of photoinhibition and the PSII repair cycle (Tyystjärvi et al., 1994). In later iterations, the parameters of the basic three-state model were augmented with a more mechanistic interpretation (Han, 2001, 2002), and the models were extended to describe the effects of intermittent light (Rubio et al., 2003). Recently, a basic three-state model was also applied to describe the kinetics of non-photochemical quenching (NPQ), induced by an orange carotenoid protein (OCP), in cyanobacteria (Gorbunov et al., 2011). To this day, simple three-state models remain relevant

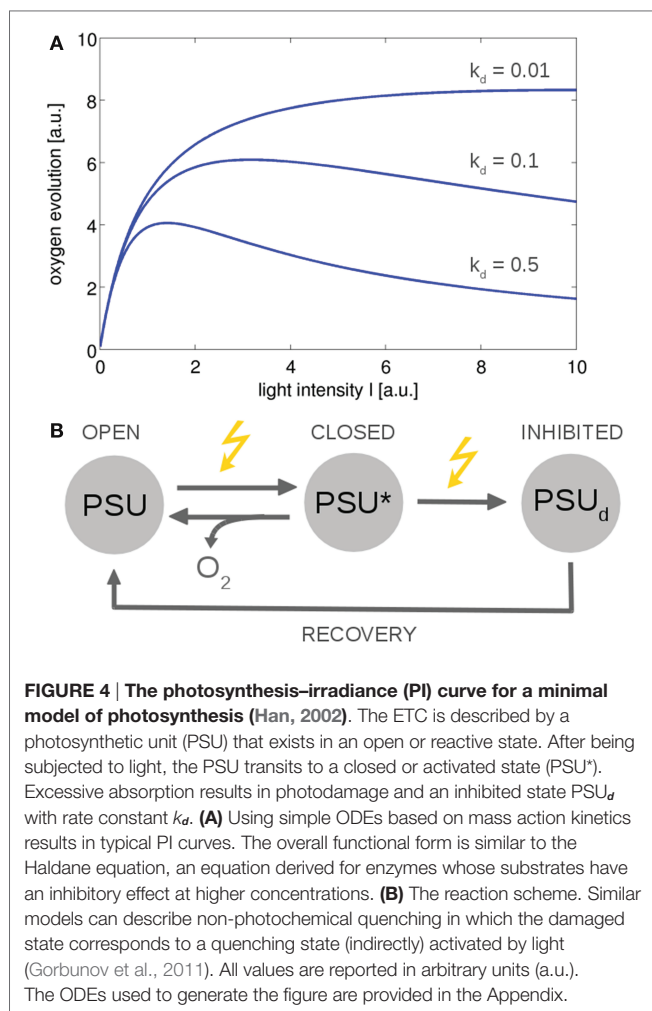


to describe overall photosynthetic activity, in particular in bulk models to assess productivity in photobioreactors and related industrial application (Nedbal et al., 2010; Bernard, 2011). With respect to understanding interactions between cellular processes, a drawback of highly simplified growth models is their insufficient representation of intracellular parameters, such as no explicit PQ pool, no explicit regeneration of ATP and NADPH, and lack of alternative electron transport.

Beyond overall bulk models of photosynthesis, there is a significant history of biophysical models to understand oxygen evolution and chlorophyll fluorescence transients, often with a focus on PSII, as well as to understand specific properties, such as energy distribution in the photosynthetic apparatus (Butler and Strasser, 1977; Butler, 1978) or, more recently, excitation transfer in the PSII membrane (Amarnath et al., 2016). Early kinetic models were described by Mar and Govindjee (1972), a more elaborate model was put forward by Holzwarth et al. (2006) and later analyzed by Nedbal et al. (2007). Further elaborate models of this kind were developed by Lazár (2003) and Zhu et al. (2005). The former was refined and extended by Jablonský and Lazár (2008); different approaches were later compared by the same authors (Lazár and Jablonský, 2009). Common to these models is a focus on chlorophyll fluorescence emission, and to

a lesser extent oxygen evolution, as the main output variables. While relevant for biophysical research, the respective models cannot be straightforwardly integrated into more comprehensive models of phototrophic growth, due to the focus on fast time scales and specific output variables. We note that the interpretation of results obtained from pulse-amplitude modulated (PAM) fluorimetry significantly differs between cyanobacteria and plants (Schuurmans et al., 2015; Acuña et al., 2016), with modeling approaches focusing almost exclusively on the latter.

Models that explicitly describe the photosynthetic electron transport chain and subsequent reactions, in addition to PSII, are more suitable to integrate into the context of a living cell. To this end, a small number of models exist (Berry and Rumberg, 2000; Vershubskii et al., 2014), typically based on ODEs. An elaborate model of this type was proposed by Laisk et al. (2006), developed to understand the photosynthetic process from light absorption to sucrose synthesis. The model neglects many of the detailed biophysical properties of earlier models (Zhu et al., 2005), such as an explicit representation of the s-states that describe the cyclic reactions of the oxygen-evolving complex (Kok et al., 1970). The model instead provides a combination of whole chain electron transport and carbon assimilation processes, including non-photochemical quenching, chlorophyll fluorescence, and (albeit



simplified) photorespiration. Using a similar approach, Zhu et al. (2013) described a detailed dynamic model of leaf photosynthesis, based on ODEs, from light capture to carbon assimilation, and incorporates the previous partial model of the same authors (Zhu et al., 2007) augmented by additional reactions of the ETC. Both models focus on C3 plant metabolism, but similar approaches are feasible for cyanobacteria.

Importantly, both models provide a sufficient level of detail to interface with other cellular processes and include ATP regeneration and reduction of NADPH, photorespiration, alternative electron transport, as well as an explicit representation of the PQ pool and lumenal pH. Owing to the focus on plant C3 metabolism, neither of the models describe peculiarities of cyanobacteria, such as shared components between the photosynthetic and respiratory ETC and the resulting differences in regulation.

Selected models of the ETC and the photosynthetic light reactions are summarized in **Table 1**. Main challenges for the development of corresponding models for cyanobacteria are to incorporate the respiratory ETC, as well as to incorporate the specific alternative electron sinks of cyanobacteria. As an interface to other cellular processes, discussed below, a model of the cyanobacterial ETC should include regulatory switches in cyanobacterial photosynthesis, regulation of light harvesting including regulation of the orange carotenoid protein, state transitions that control the relative energy transfer from phycobilisomes to PSII versus PSI, and alternative electron sinks that serve as “electron valves” and prevent overreduction of the ETC, among other features that are relevant for cyanobacterial functioning and growth (Mullineaux, 2014). Relevant exchange variables are ATP and NADPH regeneration, the state of the PQ pool, leakage of reactive oxygen species (ROS), and oxidation of metabolites for cellular respiration.

TABLE 1 | Selected models of the photosynthetic light reactions and the electron transport chain (ETC) in plants and cyanobacteria.

| | |
|----------------------------|---|
| Han (2002) | A mechanistic model of algal photoinhibition induced by photodamage to photosystem II A minimal three-state model describing the photosynthetic response to irradiance (PI) curve based on ordinary differential equations (ODEs). The model is representative of a family of similar models, including applications on non-photochemical quenching in cyanobacteria (Gorbinov et al., 2011). |
| Zhu et al. (2005) | Chlorophyll a fluorescence induction kinetics in leaves predicted from a model describing each discrete step of excitation energy and electron transfer associated with Photosystem II A detailed model of fluorescence induction in PSII in plant leaves based on ODEs. Table 3 in Zhu et al. (2005) provides a comparison of major assumptions and results of current models of fluorescence induction. |
| Laisk et al. (2006) | C3 photosynthesis <i>in silico</i> An integrated model of photosynthetic electron transport in C3 plants, including carbon reduction via the Calvin-Benson cycle based on ODEs. One of the first integrated models of the entire photosynthetic machinery. |
| Jablonský and Lazár (2008) | Evidence for intermediate S-states as initial phase in the process of oxygen-evolving complex oxidation A detailed model of PSII to describe oxygen evolution and chlorophyll fluorescence based on ODEs. |
| Jablonský et al. (2008) | Impact of dimeric organization of enzyme on its function: the case of photosynthetic water splitting An improved model of dimeric PSII describing the role of PSII dimerization on oxygen evolution. The model is based on ODEs and is an example of a very large number of variables ($>10^3$) due to the combinatorial explosion of possible states and transitions between them. |
| Ebenhöh et al. (2011) | A minimal mathematical model of non-photochemical quenching of chlorophyll fluorescence A highly simplified model to describe light harvesting and short-term adaptive quenching dynamics in the plant photosynthetic ETC based on ODEs. Variables are the three states of the reaction centers, the oxidized and reduced PQ pool, lumenal proton concentration, active and inactive quencher, as well as stromal ATP and ADP concentrations. |
| Zhu et al. (2013) | e-Photosynthesis: a comprehensive dynamic mechanistic model of C3 photosynthesis: from light capture to sucrose synthesis A detailed model of leaf photosynthesis from light harvesting to carbohydrate synthesis based on ODEs. The model is a synthesis of the earlier model of Zhu et al. (2005) and a model of the core carbon metabolism (Zhu et al., 2007). |

This table and all following tables are not exhaustive but highlight selected models that are of particular relevance to describe aspects of cyanobacterial growth. Model details, such as the exact number of variables, are listed only if they can be unambiguously sourced from the original publication.

4. KINETIC MODELS OF CELLULAR METABOLISM

The energy harvested by the photosynthetic light reactions drives the assimilation of inorganic carbon and the synthesis of storage compounds and building blocks for cellular growth. Photoautotrophic metabolism involves the uptake of inorganic carbon facilitated by CO₂-concentrating mechanisms (CCMs), assimilation of CO₂ by RuBisCO, and the subsequent synthesis of cellular building blocks mediated by a network of metabolic reactions. Computational concepts used to describe cyanobacterial metabolism have been discussed previously (Steuer et al., 2012), here we focus on the integration of such descriptions into integrative models of phototrophic growth. In particular, models of metabolism are highly diverse and span multiple orders of magnitude with respect to the time scales and number of variables involved. As highlighted previously (Steuer and Junker, 2009), no single universal computational methodology exists that is suitable to describe all relevant aspects of metabolic functioning. Rather, a hierarchy of computational approaches exists, ranging from detailed kinetic models of individual enzymatic reactions, based on ODEs, to large-scale stoichiometric reconstructions that are evaluated using constraint-based analysis.

The basic building blocks of metabolism are the actions of individual enzymes and their respective reaction mechanisms. Computational modeling of enzyme kinetics is well understood (Steuer and Junker, 2009; Sauro, 2014), even though specific reaction mechanisms and atom transition maps are not yet comprehensively available and must be confirmed on an individual per-reaction basis. Detailed kinetic models of key reactions have been proposed in the literature, most notably for RuBisCO, the key enzyme of the Calvin–Benson cycle (Witzel et al., 2010). Following the rules for overall rate equations of enzyme kinetic mechanisms, multiple reaction steps can be integrated into larger models of cellular pathways. Corresponding detailed kinetic models of metabolic pathways exist since the late 1950s, see Steuer and Junker (2009) for a review, and several kinetic pathway models have since been proposed for phototrophic plant metabolism. Of particular interest is the Calvin–Benson cycle and the adjacent carbon metabolism. Among the first computational descriptions of phototrophic C3 carbon metabolism were the models proposed by Milstein and Bremermann (1979) and Hahn (1984). The former involves 17 first-order ODEs and 22 parameters. The latter involves 19 state variables and describes the dynamics of Calvin–Benson cycle intermediates, as well as parts of sucrose and starch metabolism. The model was later extended to include photorespiration (Hahn, 1987), and simplified representations were considered (Hahn, 1991). The latter analysis demonstrated that also smaller models are able to reproduce the observed dynamics. Both models, as well as the model of Laisk and Walker (1986), are largely based on mass action kinetics, rather than derived enzyme kinetic equations for kinetic mechanisms. Parallel to the development of detailed kinetic models of core carbon metabolism, a number of biochemical models of photosynthetic CO₂ assimilation have been developed that focus on plant-specific properties, such as gas exchange and stomatal conductance, often also incorporating aspects of carbon reduction, see Farquhar et al. (1980)

for an influential early example. Likewise, a significant number of models were developed to investigate the origin of photosynthetic oscillations (Giersch, 1986; Laisk and Walker, 1986; Laisk et al., 1989; Rovers and Giersch, 1995), see Roussel and Igamberdiev (2011) for a recent review.

The prototype of most current enzyme kinetic models of the Calvin–Benson cycle was proposed by Pettersson and Ryde-Pettersson (1988). The model is based on mechanistic non-linear enzyme kinetic rate equations, implemented as ODEs, together with equilibrium mass-action ratios. The model describes the dynamics of the Calvin–Benson cycle under conditions of light and CO₂ saturation. The parameterization of the model involved approximately 50 kinetic parameters, sourced from the literature across several plant species. The model of Pettersson and Ryde-Pettersson (1988), like many current and past kinetic models, is therefore not necessarily a model of a single plant species, but must be considered as a prototype model that describes several generic aspects of plant leaf C3 metabolism. The model was later adapted to investigate further aspects of metabolic regulation in phototrophic metabolism (Poolman et al., 2000, 2001), and extended by Zhu et al. (2007) to investigate the reallocation of enzymes of photosynthetic carbon metabolism with respect to optimal nitrogen and protein investment. These models also served as a blueprint for the first detailed kinetic models of cyanobacterial core carbon metabolism. Jablonský et al. (2013) proposed a modified version of the model of Zhu et al. (2007), adapted to describe the cyanobacterium *Synechococcus elongatus* PCC 7942, to investigate the functional consequences of isoenzymes. The majority of parameters were retained from the original models. The model was later refined (Jablonský et al., 2014) to explain the metabolic regulation of primary carbon metabolism, also incorporating transcriptional data as a constraint for model dynamics. Most recently, a kinetic model of the central carbon metabolism of the cyanobacterium *Synechocystis* sp. PCC 6803 was developed to investigate the role of isozymes on metabolic network homeostasis with respect to changes in gene expression induced by different CO₂ conditions (Jablonský et al., 2016). In particular, a comparison of model properties indicated that the higher number of isozymes present in the *Synechocystis* sp. PCC 6803 genome compared to the (smaller) genome of *Synechococcus elongatus* PCC 7942 may correspond to a shift of metabolic regulatory strategies from transcriptional control in latter toward post-transcriptional control in the former (Jablonský et al., 2016).

From computational perspective, the kinetic models considered above share several features relevant to the integration into multiscale models of phototrophic growth. In each case, the dynamics of the concentrations of metabolic intermediates are described by ordinary differential equations (ODEs). Rate equations are derived from enzyme kinetic mechanisms, and implemented using (usually reversible) non-linear Michaelis–Menten type functions. The rate equations consider allosteric regulations, as well as other post-translational mechanisms, as far as such interactions are known. The light reactions are typically highly simplified. In the model of Pettersson and Ryde-Pettersson (1988) and its later adaptations, ATP is provided by a single overall reaction (an ATP synthetase) that can be modulated according to light intensity. The concentrations of NADP⁺ and

NADPH are assumed to be constant. Likewise, enzyme amounts are represented by external parameters, the respective values are part of the maximal reaction velocities V_{\max} . **Table 2** lists selected kinetic models of central carbon metabolism and the Calvin–Benson cycle. We conjecture that such models provide a reasonable account of metabolite dynamics on short and medium time scales (minutes to few hours) and to metabolic adaptations to brief periods of darkness. As yet, the construction of kinetic models to adequately represent changes in day/night metabolism remains a considerable challenge.

The key factors to integrate models of core carbon metabolism into overall models of phototrophic growth requires an explicit representation of energy (ATP) and redox state (NADPH/NADP⁺) as dynamic variables that allow coupling to the ETC. In this respect, first steps have been made for plant metabolism. The models of Laisk et al. (2006) and Zhu et al. (2013) both integrate the photosynthetic light reactions with a detailed representation of the core C3 carbon metabolism, including photorespiration. Such models provide a framework for guiding engineering efforts and allow for a description of

photosynthesis and carbon fixation in response to, for example, changes in photon flux density.

Beyond the integration of light capture and carbon metabolism, significant challenges remain. Allosteric post-translation regulation only covers small to medium time scales. No current kinetic model provides a description of diurnal changes in metabolism and the corresponding switch from carbon assimilation to the mobilization of storage compounds. In addition to redox regulation, such switches are likely to require the inclusion of additional hierarchies of cellular regulation, in particular transcription and possibly regulation by the circadian clock. Switches in metabolic modes are of particular relevance for cyanobacterial metabolism and growth, as they lack the compartmentation of eukaryotic algae and plants. Likewise, current models focus on carbon metabolism and its regulation, limitation of other macronutrients like phosphorus or nitrogen is typically not considered. Nonetheless, in particular with respect to nitrogen, kinetic models can be expected to provide insight into the role of certain metabolites, such as 2-oxoglutarate (2-OG), as signaling compounds (Fokina et al., 2010).

TABLE 2 | Selected kinetic models of the central carbon metabolism in photosynthetic organisms.

| | |
|---------------------------------------|---|
| Farquhar et al. (1980) | A biochemical model of photosynthetic CO₂ assimilation in leaves of C3 species An influential early model to describe photosynthetic CO ₂ assimilation <i>in vivo</i> . The model describes limiting processes in the leaf and integrates these into overall systems behavior. |
| Hahn (1984) | A mathematical model of leaf carbon metabolism Among the first kinetic models of plant leaf carbon metabolism. The model is based on ODEs for 19 metabolic intermediates using mass action kinetics. |
| Hahn (1987) | A mathematical model of photorespiration and photosynthesis An extension of the earlier ODE model of Hahn (1984) that incorporates photorespiration (glycolate and glycerate pathways) and diffusion between atmosphere and mesophyll tissue using 31 variables. |
| Pettersson and Ryde-Pettersson (1988) | A mathematical model of the Calvin photosynthesis cycle The prototype of most current kinetic models of plant leaf metabolism. The model is based on ODEs and makes use of derived enzyme kinetic rate equations for 9 reactions. 11 reactions are assumed to be close to equilibrium and are represented by algebraic equations within the model. Parameters are sourced from the biochemical literature. |
| Laisk et al. (1989) | A mathematical model of the carbon metabolism in photosynthesis. Difficulties in explaining oscillations by fructose 2,6-bisphosphate regulation The model is similar in scope to the model of Pettersson and Ryde-Pettersson (1988) and provides a detailed description of core carbon metabolism (the Calvin–Benson cycle, starch and sucrose synthesis) involving 20 ODEs. The light reactions are described by a single enzyme-catalyzed reaction, the ATP synthetase. The NADPH/NADP ⁺ ratio is assumed to be constant. |
| Poolman et al. (2000) | Modeling photosynthesis and its control An adaptation of the model of Pettersson and Ryde-Pettersson (1988). The major difference is that reactions close to equilibrium are represented by reversible mass action kinetics rather than algebraic equations. The model was later analyzed with respect to its control properties and shows evidence for two steady states (Poolman et al., 2001). |
| Zhu et al. (2007) | Optimizing the distribution of resources between enzymes of carbon metabolism can dramatically increase photosynthetic rate: a numerical simulation using an evolutionary algorithm An extended kinetic model of plant leaf metabolism based on the model of Pettersson and Ryde-Pettersson (1988) including a detailed representation of photorespiratory metabolism. The model was used to study the optimal partitioning of enzymes to increase photosynthetic rate. |
| Jablonský et al. (2013) | Phosphoglycerate mutases function as reverse regulated isoenzymes in <i>Synechococcus elongatus</i> PCC 7942 A modified and corrected model for cyanobacterial central carbon metabolism, based on the models of Pettersson and Ryde-Pettersson (1988) and Zhu et al. (2007). The model is the first kinetic model specific for a cyanobacterium and was used to investigate the function of phosphoglycerate mutase isoforms in <i>Synechococcus elongatus</i> PCC 7942. The model was later improved and extended (Jablonský et al., 2014) to investigate the diverse roles of isoenzymes in the same strain. |
| Jablonský et al. (2016) | Different strategies of metabolic regulation in cyanobacteria: from transcriptional to biochemical control A model of the central carbon metabolism of <i>Synechocystis</i> sp. PCC 6803 based on the earlier model of Jablonský et al. (2014). The model describes the dynamics of 36 metabolite concentrations interconnected by 54 reactions and makes use of 182 kinetic parameters. |

The models of Laisk et al. (2006) and Zhu et al. (2013) integrate the photosynthetic light reactions and the core carbon metabolism and are already listed in **Table 1**. Minimal models that investigate oscillatory mechanisms, and models that primarily focus on plant-specific properties, such as stomatal conductance, are not considered.

A major obstacle for detailed kinetic models also remains the scarcity of enzyme kinetic data. The construction of kinetic models requires detailed knowledge of enzyme kinetic parameters—with typically 4–5 parameters per reaction, including the Michaelis–Menten parameters K_M for substrates and products, as well as thermodynamic equilibrium values K_{eq} and the specific catalytic activities of enzymes. Even metabolic pathways of moderate size, such as the Calvin–Benson cycle and adjacent reaction, typically consist of 20–30 enzymatic reactions. Therefore, the construction of larger kinetic models, while feasible from a computational point of view, is primarily limited by data availability and data reliability (Srinivasan et al., 2015). To some extent, the scarcity of information about kinetic parameters can be alleviated by explicitly accounting for uncertainty in kinetic models of metabolism—suitable approaches have been proposed recently (Wang et al., 2004; Steuer et al., 2006; Tran et al., 2008; Steuer and Junker, 2009; Murabito et al., 2014) but are not yet widely applied in models of phototrophic growth.

5. MODELS OF CARBON-CONCENTRATING MECHANISMS

A characteristic feature of cyanobacterial growth is the use of CO_2 -concentrating mechanisms (CCMs) to facilitate the uptake and acquisition of inorganic carbon. CCMs allow cyanobacteria to raise the local concentration of CO_2 in the vicinity of the carboxylating enzyme RuBisCO, and thereby overcome the comparatively low affinity of RuBisCO for CO_2

and depress the competitive oxygenation reaction (photorespiration). Cyanobacterial CCMs typically make use of dedicated microcompartments, the carboxysomes, that separate the assimilation of CO_2 by RuBisCO from the rest of the cell. The CCMs of cyanobacteria relies on a number of components. The respective mechanisms are reasonably well understood (Kaplan and Reinhold, 1999; Price et al., 2008; Burnap et al., 2015), see **Figure 5** for a schematic depiction, and the requirement of a quantitative mathematical analysis has recently been highlighted (Mangan and Brenner, 2014). The efficiency of the cyanobacterial CCMs can be characterized by the ratio between the apparent whole-cell affinity for extracellular CO_2 and the respective affinity for CO_2 of the carboxylating enzyme RuBisCO—with ratios up to 1,000 reported in the literature (Burnap et al., 2015). While many components of the CCMs are constitutively expressed, the expression of specific uptake systems is differentially regulated depending on environmental parameters, in particular light intensity and the availability of inorganic carbon (Ci) (Kaplan and Reinhold, 1999; Burnap et al., 2015).

In addition to their important function to enhance the local CO_2 concentration and depressing photorespiration, CCMs may also be involved in dissipating excess light energy (Xu et al., 2008) and might play a role in pH homeostasis. Despite this tight integration with carbon metabolism, none of the current kinetic models of carbon metabolism explicitly accounts for CCMs. Nonetheless, a number of quantitative models of CCMs are available that can be readily incorporated into models of cyanobacterial growth. See **Table 3** for selected computational models of CCMs. Early

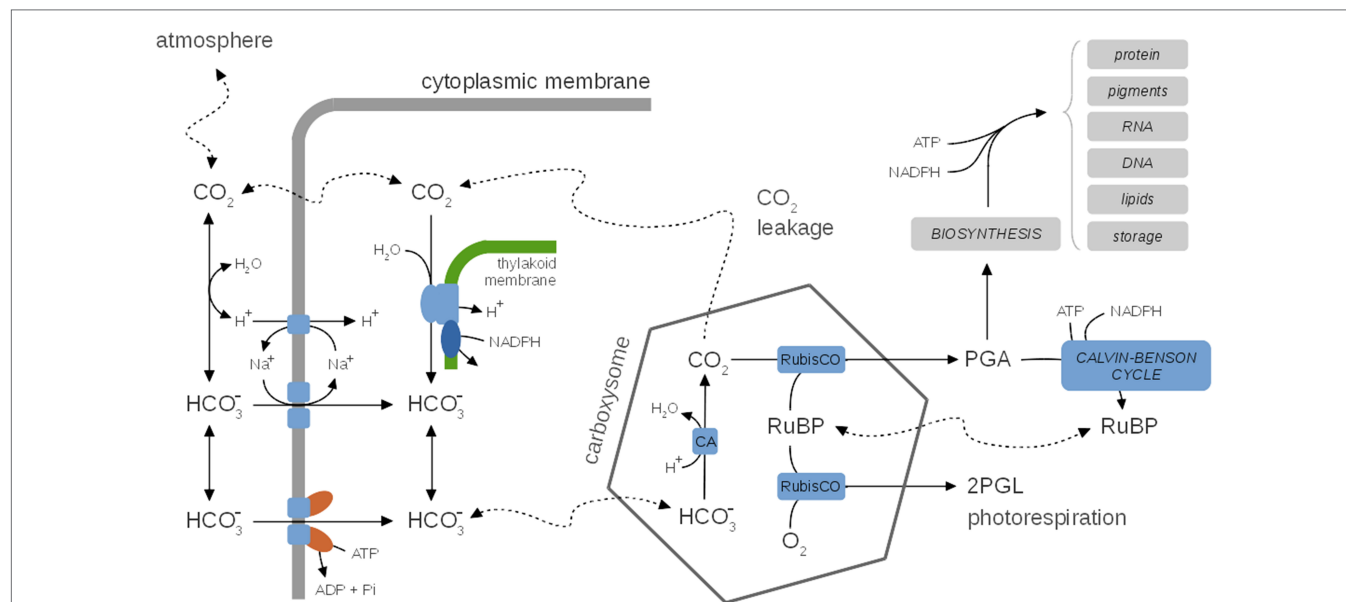


FIGURE 5 | A generic representation of cyanobacterial CO_2 -concentrating mechanisms (CCM), species-specific differences are neglected. Inside, the carboxysome bicarbonate (HCO_3^-) is rapidly converted to CO_2 via the carbonic anhydrase (CA). CCMs thereby facilitate a CO_2 gradient to increase the local concentration of CO_2 in the vicinity of RuBisCO and depress photorespiration. Several uptake mechanisms for HCO_3^- exist, including ABC-type high affinity transporters and $\text{Na}^+/\text{HCO}_3^-$ symporters. The activity of CCMs, including expression of several components, is modulated by environmental parameters, in particular CO_2 availability. CO_2 leaking from the carboxysomes may diffuse into the medium or is partly converted back to HCO_3^- (carbon cycling) at the thylakoid membrane using insufficiently understood mechanisms. CCMs utilize cellular energy and might be involved in dissipating excess light energy and play a role in the maintenance of internal pH. Hence, CCMs are integrally tied to cellular metabolism and growth. See **Table 3** for selected computational models of CCMs.

TABLE 3 | Selected models of cyanobacterial CO₂-concentrating mechanisms (CCMs).

| | |
|---------------------------|---|
| Badger et al. (1985) | A model for HCO₃⁻ accumulation and photosynthesis in the cyanobacterium <i>Synechococcus</i> sp: theoretical predictions and experimental observations A simple early model of CCMs in cyanobacteria based on a single spherical compartment into which inorganic carbon is actively accumulated. |
| Fridlyand et al. (1996) | Quantitative evaluation of the role of a putative CO₂-scavenging entity in the cyanobacterial CO₂-concentrating mechanism Based on a series of earlier models (Reinhold et al., 1991), the model provides a quantitative account of cyanobacterial CCM with a focus on the implications of scavenging CO ₂ that leaks outwards from the carboxysome. The model is based on eight differential equations that describe diffusion and transport of inorganic carbon (CO ₂ and HCO ₃ ⁻) between the various compartments. |
| Hopkinson et al. (2014) | The minimal CO₂-concentrating mechanism of <i>Prochlorococcus</i> spp. MED4 is effective and efficient A simple model based on the model of Reinhold et al. (1987) and adapted from <i>Prochlorococcus</i> spp. MED4. Transfer of inorganic carbon is described by four differential equations. |
| Clark et al. (2014) | Insights into the industrial growth of cyanobacteria from a model of the carbon-concentrating mechanism A kinetic model of the CCM parameterized for two cyanobacterial species and a hypothetical no-CCM mutant. The work assumes the carboxysome shell to be impermeable to CO ₂ and concludes that carboxysome geometry is unimportant and interspecies differences in CCMs are largely due to active HCO ₃ ⁻ transporters. |
| Mangan and Brenner (2014) | Systems analysis of the CO₂-concentrating mechanism in cyanobacteria A quantitative evaluation of cyanobacterial CCMs based on a reaction–diffusion model to investigate parameter ranges that give rise to a functional and effective CCM. |
| Mangan et al. (2016) | pH determines the energetic efficiency of the cyanobacterial CO₂-concentrating mechanism An update of the model of (Mangan and Brenner, 2014) that introduces intracellular pH as a key physiological parameter that determines the energetic costs associated with CCMs and carbon fixation. |

Modeling the CCM typically involves a spatial component. Analysis of the respective partial-differential equations (PDEs) shows, however, that intracompartamental concentration gradients can be neglected. Models of the CCM are therefore typically based on ODEs, rather than PDEs.

models focus either on simple equations for CO₂ and HCO₃⁻ (bicarbonate) flux into and out of the cell (Badger et al., 1985), as well as on arguments based on reaction–diffusion equations (Reinhold et al., 1987, 1991). These models were refined further to include explicit representations of the carboxysomes. Specifically, the model of Fridlyand et al. (1996) considers the various CO₂ or HCO₃⁻ fluxes between medium, periplasmic space, cytoplasm and carboxysomes using derived values for geometric parameters, and permeability and diffusion coefficients. The model also considers the energetic consequences of scavenging CO₂ that leaks back into the cytoplasm. Models can be adopted to specific cyanobacterial strains, such as the model of Hopkinson et al. (2014) for *Prochlorococcus* spp. MED4. Two recent quantitative models of CCM functioning are available (Clark et al., 2014; Mangan and Brenner, 2014). Both models are based on reaction–diffusion equations that are solved for highly simplified spatial topologies (spherical cells), but otherwise make use of partially divergent assumptions. The focus of Clark et al., 2014 are interspecies differences and a hypothetical carboxysome-free mutant that is of interest in industrial settings with elevated CO₂ supply. The model assumes that the carboxysome walls are impermeable to CO₂. The model of Mangan and Brenner (2014) assumed that carboxysome permeability is identical for HCO₃⁻ and CO₂ and the model explores the range of best parameter values that give rise to a functional and effective CCM. While carboxysome permeability has not yet been measured directly, Mangan and Brenner (2014) concluded that optimal parameter values indeed exist, and transport rates and concentrations derived for these optimal values are in good agreement with known experimental data. Very recently, the model was extended to incorporate the effect of intracellular pH as a key physiological parameter that governs the composition of the Ci pool (Mangan et al., 2016). The “pH-aware” model highlights the utility of quantitative models to evaluate the energetic costs of Ci accumulation for CCMs.

While current models of the CCM do consider the optimality and functioning of CCMs under different intracellular and environmental conditions, they typically do not incorporate explicit models of cellular growth or other cellular mechanisms. However, the high energy demand of Ci transport, either by direct hydrolysis of one ATP per bicarbonate, or indirectly via the costs of ion transport, the costs of synthesizing carboxysome shell proteins, as well as the significant impact of CCMs on the efficiency of carbon assimilation suggest that further integration of models of CCMs into a broader context of cellular functioning is worthwhile to understand the trade-offs and interactions between energy investment, CCM utilization, carbon assimilation, and growth.

6. LARGE-SCALE MODELS OF CYANOBACTERIAL METABOLISM

Beyond kinetic models of central carbon metabolism, metabolic networks are increasingly described using large-scale stoichiometric reconstructions (Steuer et al., 2012). Metabolic reconstructions aim to provide a comprehensive account of all possible interconversions of small molecules within a given cell or compartment, including enzymatic reactions as well as non-catalyzed (spontaneous) interconversions, transport reactions, and diffusion. Metabolic reconstructions are derived from annotated genomes with subsequent steps of manual curation and gap filling, see Knoop et al. (2010), Steuer et al. (2012), and Knoop et al. (2013) for recent examples. The description typically involves only knowledge about the stoichiometry of interconversions; knowledge about kinetic parameters (such as Michaelis–Menten parameters) or allosteric regulation is not required.

Nonetheless, large-scale stoichiometric models of bacterial metabolism are highly predictive (McCloskey et al., 2013). The predictive power derives from the fact that the fluxes through

enzymatic reactions are not independent. Constraint-based computational methods rely on the fact that under steady-state conditions most intracellular metabolites do not accumulate. The rate of synthesis of any non-accumulating metabolite must therefore approximately equal the rate of consumption of this metabolite. Similar arguments also hold for diurnal metabolism: if, after a full diurnal cycle, the concentration of a given intracellular metabolite is approximately equal to its initial value, then the total flux of synthesis reactions and the total flux of consuming reactions must be approximately equal. The condition of flux balance puts significant constraints on the feasible flux space. Predictions about specific flux solutions are then typically based on assumptions about metabolic optimality. That is, among all feasible flux solutions, constraint-based methods seek to identify a solution that maximizes a given objective function, such as the maximal yield of biomass for a given light input—motivated by the fact that a similar selection might take place during evolution. Predictions from large-scale stoichiometric models are therefore not mechanistic, that is, they are not derived from knowledge about biophysical or biochemical interactions. Rather, predictions are derived from how metabolism *ought* to behave given the assumption that metabolic functioning fulfills certain evolutionary optimality principles.

Computationally, the analysis of large-scale stoichiometric reconstructions is based on methods of linear programming (LP) and is computationally feasible also for networks consisting of several thousands of reactions. The strength of stoichiometric models and constraint-based analysis are questions such as the following: What is the maximal growth yield for a given light or carbon input? Which set of enzymes is essential for the synthesis

of certain biomass components? How many distinct biochemical paths exist for the synthesis of certain biomass components and how do these pathways differ with respect to cellular energy expenditure and cofactor utilization? Due to the specific computational methodology, however, a direct integration of large-scale stoichiometric models into kinetic models of metabolism remains challenging. Various extensions toward incorporating dynamics have been proposed (Mahadevan et al., 2002; Kim et al., 2008; Feng et al., 2012; Antoniewicz, 2013), and extensive efforts are undertaken to bridge the gap between kinetic and stoichiometric models (Steuer, 2007; Steuer and Junker, 2009; Chakrabarti et al., 2013; Srinivasan et al., 2015).

Detailed stoichiometric reconstructions are available for several cyanobacterial strains (Knoop et al., 2010, 2013; Montagud et al., 2010; Hamilton and Reed, 2012; Nogales et al., 2012; Saha et al., 2012; Vu et al., 2012; Mueller et al., 2013; Maarleveld et al., 2014; Yoshikawa et al., 2015)—typically consisting of several hundred enzymatic interconversions and accounting for all known pathways related to central metabolism and the synthesis of key biomass components. See Table 4 for selected examples of metabolic reconstructions. Such large-scale reconstructions are valuable tools to derive consistent equations for the (maximal) growth yield with respect to light input, to derive core models of reaction pathways, and to make predictions about maximal product yield in biotechnological applications (Zavřel et al., 2016). In particular, large-scale reconstructions also enable a semiautomated extraction of meaningful core models to facilitate the construction of smaller kinetic models (Erdreich et al., 2015). Analysis of the respective networks, however, is often confined to either a constant light environment, or to heterotrophic growth

TABLE 4 | Selected large-scale metabolic reconstructions of cyanobacteria.

| | |
|---------------------------|---|
| Shastri and Morgan (2005) | Flux balance analysis of photoautotrophic metabolism One of the first metabolic reconstruction and stoichiometric evaluations of cyanobacterial central carbon metabolism. The model is medium scale with a focus on central metabolism and does not include gene–reaction associations. |
| Knoop et al. (2010) | The metabolic network of <i>Synechocystis</i> sp. PCC 6803: systemic properties of autotrophic growth A large-scale (but not genome-scale) reconstruction of <i>Synechocystis</i> sp. PCC 6803 including gene–reaction associations and photorespiration. The model includes the GABA shunt to close the (incomplete) TCA cycle. |
| Montagud et al. (2010) | Reconstruction and analysis of genome-scale metabolic model of a photosynthetic bacterium A reconstruction of <i>Synechocystis</i> sp. PCC 6803. The model also includes a putative glyoxylate shunt that subsequently could not be identified <i>in vivo</i> . |
| Saha et al. (2012) | Reconstruction and comparison of the metabolic potential of cyanobacteria <i>Cyanothece</i> sp. ATCC 51142 and <i>Synechocystis</i> sp. PCC 6803 A detailed reconstruction of the cyanobacterium <i>Cyanothece</i> sp. ATCC 51142. The model considers separate (light/dark) biomass equations to reflect the differences between light and dark phases. |
| Vu et al. (2012) | Genome-scale modeling of light-driven reductant partitioning and carbon fluxes in diazotrophic unicellular cyanobacterium <i>Cyanothece</i> sp. ATCC 51142 A reconstruction of <i>Cyanothece</i> sp. ATCC 51142 with a focus on relative fluxes through the ETC. |
| Nogales et al. (2012) | Detailing the optimality of photosynthesis in cyanobacteria through systems biology analysis An expanded reconstruction of <i>Synechocystis</i> sp. PCC 6803 including gene–reaction associations. The model does not yet incorporate the TCA bypass identified by Zhang and Bryant (2011). |
| Knoop et al. (2013) | Flux balance analysis of cyanobacterial metabolism: the metabolic network of <i>Synechocystis</i> sp. PCC 6803 An expanded reconstruction of <i>Synechocystis</i> sp. PCC 6803, including the TCA bypass (Zhang and Bryant, 2011) as well as the original GABA shunt. The model is backed by experimental analysis to identify or disprove the existence of a functional glyoxylate shunt and considers diurnal changes in reaction fluxes. |
| Mueller et al. (2013) | Rapid construction of metabolic models for a family of cyanobacteria using a multiple source annotation workflow Genome-scale reconstructions of several <i>Cyanothece</i> strains that demonstrate the advantages of using a parallel workflow. |

Reconstructions of algae and land plants are not considered.

on extracellular carbon sources. Only recently stoichiometric analysis of phototrophic metabolism explicitly described different phases of light availability. For example, Knoop et al. (2013) simulated biomass synthesis fluxes over a full diurnal cycle, Muthuraj et al. (2013) used dynamic FBA (dFBA) to capture light-dark metabolism over discretized time intervals, Knies et al. (2015) considered storage metabolites that accumulate and are consumed over a diurnal cycle using a reconstruction of the unicellular alga *Emiliania huxleyi*, Cheung et al. (2014) described a flux balance model that captures interactions between light and dark metabolism in C3 and CAM leaves, and Baroukh et al. (2014) proposed a novel dynamic modeling framework to describe carbon metabolism of unicellular microalgae.

Beyond conventional FBA and related constraint-based methods, there has been increasing interest to evaluate cellular metabolism in terms of a cellular “protein economy” (Molenaar et al., 2009) and to study trade-offs in cellular resource allocation (Goel et al., 2012; Müller et al., 2015)—a theme where, as noted by Schaechter (2015), the bacterial growth physiology of old is connected to the systems biology of today (Stouthamer, 1973; Neidhardt et al., 1990). As one of the first applications involving cyanobacteria, Burnap (2015) formulated a model of autotrophic growth in terms of allocating protein resources among core functional groups, such as the ETC, light-harvesting antennae, and ribosomes. Along similar lines, Rügen et al. (2015) formulated a self-consistent autocatalytic model of phototrophic growth. The model is based on the observation that the macromolecules that constrain cellular growth, including the components of the ETC, metabolic enzymes, and ribosomes, are itself products of metabolism. Phototrophic growth can therefore be formulated as a time-dependent linear optimization problem, such that optimal growth entails a time-dependent allocation of resources during a full diurnal cycle. The approach of Rügen et al. (2015), denoted as conditional FBA, results in dynamic time courses for all involved reaction fluxes, as well as changes in biomass composition over a diurnal cycle. Similar to conventional FBA, models of this kind are not based on mechanistic insight, but rather seek to evaluate the optimality of resource allocation during phototrophic growth. It is expected that methods and applications that go beyond conventional FBA and involve spatial and temporal metabolic modeling based on genome-scale reconstructions of microbial metabolism will play an increasingly important role (Henson, 2015).

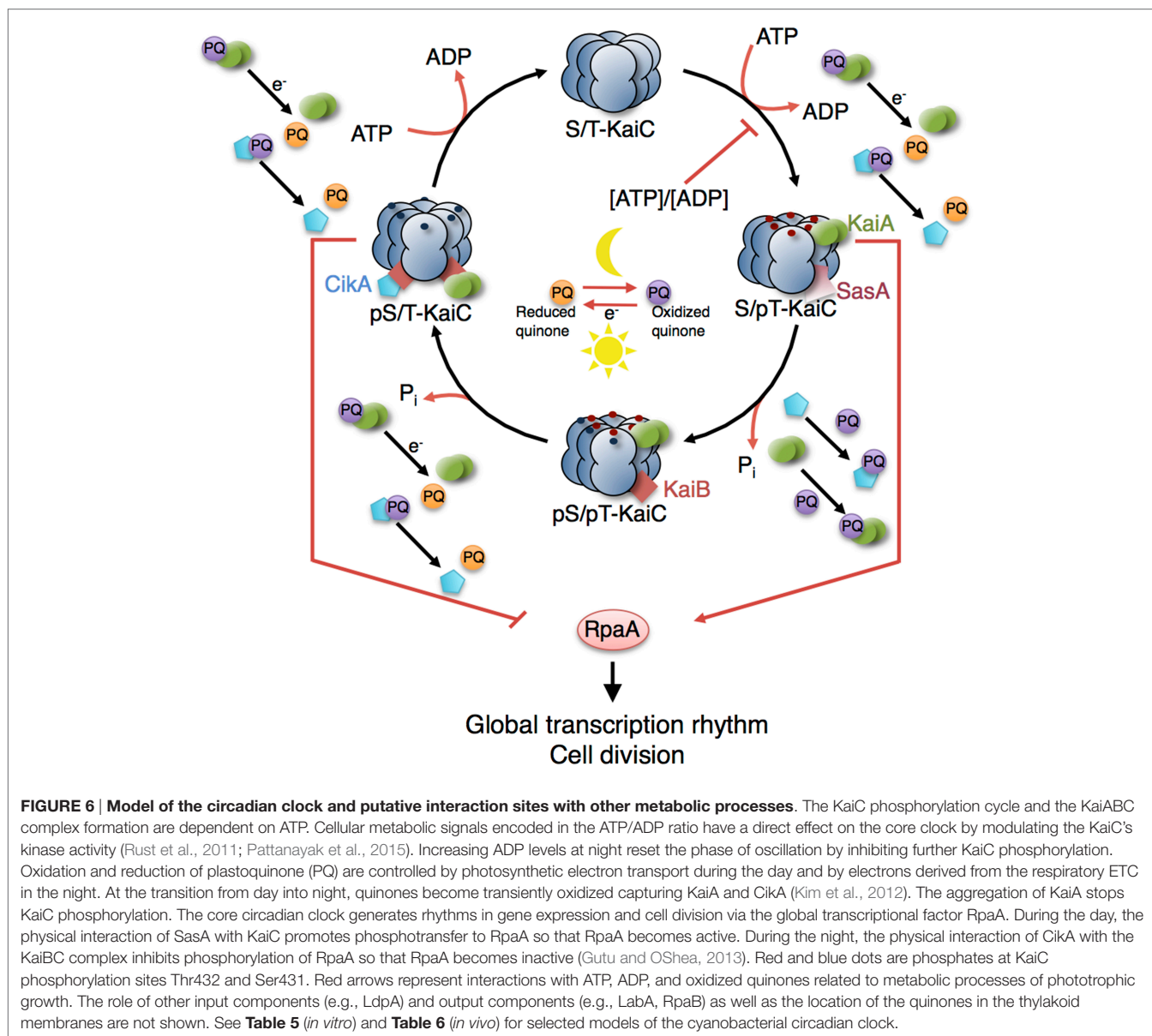
7. MODELS OF THE CYANOBACTERIAL CLOCK

In addition to the biochemistry of metabolism, phototrophic growth is highly dependent on regulatory networks to coordinate growth and to relay environmental information. To this end, of particular relevance is the cyanobacterial circadian clock—discovered in the late 1980s and unique among prokaryotes (Pattanayak and Rust, 2014). The cyanobacterial clock consists of an interrelated network of multifunctional components functioning in timekeeping, input and/or output mechanisms. *Synechococcus elongatus* PCC 7942 is the cyanobacterium whose clock is currently best studied. The core oscillator comprises only

three proteins: KaiA, KaiB, and KaiC (Ishiura et al., 1998). KaiC exhibits an intrinsic kinase, dephosphorylation, and ATPase activity (Nishiwaki et al., 2004; Terauchi et al., 2007; Egli et al., 2012; Nishiwaki and Kondo, 2012). In complex with KaiA and KaiB, KaiC undergoes circadian Thr/Ser phosphorylation (Nakajima et al., 2005; Nishiwaki et al., 2007; Rust et al., 2007) (**Figure 6**). KaiA promotes and KaiB represses phosphorylation of KaiC (Iwasaki et al., 2002; Kitayama et al., 2003; Xu et al., 2003). The ATPase activity of KaiC is extremely weak (only 15 ATP molecules are consumed per day) and slow, determining the about 24-h period of the clock (Terauchi et al., 2007; Abe et al., 2015; Chang et al., 2015). The circadian rhythm of KaiC phosphorylation runs without transcription–translation and can even operate in a test tube (Nakajima et al., 2005; Tomita et al., 2005). *In vivo*, the KaiABC protein system works as a post-translational oscillator (PTO).

The KaiC phosphorylation rhythm has widely been studied in systems biology, and hence, a variety of mathematical models have been put forward. The first models were rather minimal to explain how sustained oscillations in phosphorylation of KaiC occur, including no intermediate steps of phosphorylation, introducing feedbacks on KaiC phosphorylation or assuming hypothetical states of KaiA, KaiB, and KaiC (Emberly and Wingreen, 2006; Kurosawa et al., 2006; Mehra et al., 2006; Axmann et al., 2007; Mori et al., 2007). Emberly and Wingreen (2006) were the first who showed theoretically that monomer shuffling between KaiC hexamers at specific clock times could explain the robustness and resilience of the circadian clock—a hypothesis stated prior to experimental evidence. Different variations of the concept, monomer shuffling, have afterward been modeled by other groups (Kageyama et al., 2006; Mehra et al., 2006; Mori et al., 2007; Yoda et al., 2007; Eguchi et al., 2008; Nagai et al., 2010). Another hypothesis of how synchrony within the Kai oscillator could be achieved stresses KaiA sequestration into KaiABC complexes (Kurosawa et al., 2006; Clodong et al., 2007; Rust et al., 2007; van Zon et al., 2007; Brettschneider et al., 2010). The consensus view has emerged that both mechanisms work in concert (Qin et al., 2010a). See **Table 5** (*in vitro*) and **Table 6** (*in vivo*) for selected models of the cyanobacterial circadian clock.

Recent studies have increased our understanding of how the core oscillator is integrated with input pathways and output pathways, which enable the clock to synchronize (“entrain”) to the 24-h period of the environment and to transmit temporal information to downstream processes resulting in circadian rhythms in cellular physiology. Clock input cues involve the cellular ATP/ADP ratio, which has a direct effect on the core clock by modulating the KaiCs kinase activity (Rust et al., 2011). In particular, an increase in the ADP levels, occurring when cells are placed into darkness, inhibits further KaiC phosphorylation and thus resets the phase of oscillation to synchronize to the metabolic state of the cell. By simulating various ATP/ADP ratios that mimic different night phases, Rust et al. (2011) could recreate phase shifts in the core oscillator as seen *in vitro* and *in vivo*. For their theoretical analysis, the authors modified a previous mathematical model of the circadian clock, which was based on the rates of phosphorylation and dephosphorylation at Thr432 and Ser431 (Rust et al., 2007). In the refined model, the



KaiA-dependent kinase rates were now additionally modulated by the ratio of ATP to ATP + ADP. This model was again adapted to account for an additional ATPase activity experimentally found in the CII subunit of KaiC and required for binding of KaiB to Ser-phosphorylated KaiC (Phong et al., 2013). The KaiBC complex formation was shown to depend on an ATPase, but whose activity was insensitive to changes in the cellular ATP/ADP ratio—in contrast to the ATPase in the CII subunit of KaiC (responsible for the Thr/Ser phosphorylation reactions). The results of the combined modeling and experimental study (Phong et al., 2013) suggest that these two differently sensitive catalytic domains are responsible for the capability of the clock to receive input signals while preserving circadian rhythmicity. Depending on the specific question under investigation, both entrainment models could relatively straightforwardly be integrated with other modules of cyanobacterial physiology,

with ATP as key to coupling (e.g., light reactions). Another model version already exists, which additionally accounts for the transcription and translation of clock genes as well as the feedback to the core oscillator but needs to be extended to include interactions with other cellular parameters such as the ATP/ADP ratio (Teng et al., 2013). The mathematical clock model proposed by Brettschneider et al. (2010) could equally be envisaged for the integration into larger models of phototrophic growth. The core oscillator is modeled by a larger set of ODEs (12 ODEs; for comparison, 3 ODEs in Rust et al. (2007)) but includes hexamerization of KaiC, monomer shuffling, and assembly and disassembly of KaiAC, KaiBC, and KaiABC complexes. Here also, an extended model (15 ODEs) coupled to a transcription–translation circuit has been proposed (Hertel et al., 2013). Inhibition of KaiCs' kinase activity by ADP can be incorporated into the model.

TABLE 5 | Selected *in vitro* models of the cyanobacterial circadian clock.

| | |
|--|---|
| Emberly and Wingreen (2006) | Hourglass model for a protein-based circadian oscillator An early minimal model to explain sustained oscillations in phosphorylated KaiC. The authors suggest monomer exchange between KaiC hexamers to contribute to the robustness of the clock, prior to experimental evidence of such mechanisms. The model consists of 5 variables and 8 parameters. |
| Mehra et al. (2006) | Circadian rhythmicity by autocatalysis An early minimal model using a two-state approximation (KaiC exists in two phosphorylation states, high and low). Oscillations emerge due to autocatalytic KaiA–KaiC interaction. The model consists of eight variables and 10 parameters. |
| Mori et al. (2007) | Elucidating the ticking of an <i>in vitro</i> circadian clockwork The model describes explicitly the kinetics of complex formation and dissociation of KaiC hexamers with k phosphorylation sites per monomer. Monomer exchange in phosphorylation and dephosphorylation phase explains robust oscillations. The model uses a stochastic matrix model for simulation of hexamer kinetics resulting in a combinatorial number of variables (4,374 variables for $k = 2$ parameters). |
| Yoda et al. (2007) | Monomer-shuffling and allosteric transition in KaiC circadian oscillation Similar to Mori et al. (2007) but using deterministic equations to simulate the dynamics (56 variables, 34 parameters). |
| van Zon et al. (2007) | An allosteric model of circadian KaiC phosphorylation The model uses 2 phosphorylation states for KaiC, KaiA exhibits different binding affinity for different phospho-forms of KaiC. The model emphasizes sequestration of KaiA in KaiA–KaiB–KaiC complexes. The model consists of 30 variables and 17 parameters. |
| Clodong et al. (2007) | Functioning and robustness of a bacterial circadian clock The minimal model was designed to examine different possible feedback mechanisms to determine what type of feedback generates the most robust rhythms. The ODE model consists of 14 variables and 19 parameters. |
| Rust et al. (2007) | Ordered phosphorylation governs oscillation of a three-protein circadian clock An early ODE model to describes the dynamics of the 4 phosphorylation states of KaiC during a circadian cycle. KaiA is sequestered by serine phosphorylated KaiC using phenomenological assumptions. The model consists of 3 variables and 13 parameters. |
| Axmann et al. (2007) | A minimal circadian clock model A minimal ODE clock model using a two-state approximation of KaiC. KaiA-dependent phosphorylation is highly non-linear, sequestration of free KaiA by phosphorylated KaiC–KaiB complex favors dephosphorylation. The model was designed to study robustness of oscillations with respect to concerted changes in Kai protein concentration but fails to be invariant against such changes (6 variables, 9 parameters). |
| Eguchi et al. (2008) | Mechanism of robust circadian oscillation of KaiC phosphorylation <i>in vitro</i> A similar model as Yoda et al. (2007) and Mori et al. (2007). Focus on monomer shuffling to synchronize oscillations. The model consists of 14 variables and 4 parameters. |
| Nagai et al. (2010) | Synchronization of circadian oscillation of phosphorylation level of KaiC <i>in vitro</i> Similar to Mori et al. (2007) with focus on monomer shuffling. The model incorporates collective shifts from tense to relaxed states (14 variables and 11 parameters). |
| Brettschneider et al. (2010) | A sequestration feedback determines dynamics and temperature entrainment of the KaiABC circadian clock An ODE model to describe the dynamics of the four phosphorylation states of KaiC during a circadian cycle including complex formation and monomer shuffling (12 variables, 26 parameters). |
| Qin et al. (2010a) | Intermolecular associations determine the dynamics of the circadian KaiABC oscillator An elaborate model to describe complex formation dynamics. The model consists of 56 variables and 34 parameters. |
| <i>In vitro</i> models including energy status | |
| Rust et al. (2011) | Light-driven changes in energy metabolism directly entrain the cyanobacterial circadian oscillator A modified version of the model of Rust et al. (2007). The model includes inhibition of kinase activity by ADP. The ATP/ADP ratio is an explicit parameter within the model (3 variables, 16 parameters). |
| Phong et al. (2013) | Robust and tunable circadian rhythms from differentially sensitive catalytic domains A modified version of the model of Rust et al. (2011) with another ATPase activity that is insensitive to changes in ATP/ADP. The ATP/ADP ratio is an explicit parameter within the model (6 variables, 25 parameters). |

The models focus on the functioning of the post-translational oscillator (PTO). ATP, NADPH, and redox state are typically not considered as explicit variables. All models are implemented as ODEs, except otherwise noted.

Sensing the decrease in the ATP/ADP ratio during night is assumed to allow the clock to infer the length of night (Rust et al., 2011). In two biochemical studies, Pattanayak et al. uncovered a further connection between the clock and cellular metabolism: metabolic rhythms produced by the clock (such as rhythms in glycogen abundance, which go along with changing levels of ATP/ADP) feed back to the core oscillator. These rhythms are very likely the main driving force of the clock, allowing the cells to anticipate the onset of darkness in advance (Pattanayak et al.,

2014, 2015). In addition, the clock seems to be able to anticipate nightfall through the plastoquinone pool, which is part of the ETC. In particular, the plastoquinone pool embedded within the thylakoid membrane becomes transiently oxidized at the transition from day into night and binding of PQ to KaiA causing aggregation and decay that, in turn, reduces the positive effect of KaiA on KaiC phosphorylation (Wood et al., 2010; Kim et al., 2012). Other redox-sensitive input components such as CikA (circadian input kinase) and LdpA (light-dependent period

TABLE 6 | Selected *in vivo* models of the cyanobacterial circadian clock.

| | |
|------------------------|--|
| Kurosawa et al. (2006) | A model for the circadian rhythm of cyanobacteria that maintains oscillation without gene expression The model is based on a two-state approximation and sequestration of KaiA sequestration as mean of synchronization. The model consists of 6 variables and 18 parameters. |
| Miyoshi et al. (2007) | A mathematical model for the Kai protein-based chemical oscillator and clock gene expression rhythms in cyanobacteria The models couples the TTFL and PTO, fully phosphorylated KaiC promotes <i>kaiBC</i> transcription. The model consists of 13 variables and 32 parameters and assumes hypothetical states (not seen in experiments). |
| Zwicker et al. (2010) | Robust circadian clocks from coupled protein modification and transcription–translation cycles A PTO and coupled TTFL–PTO model. The model consists of 30 variables and 37 parameters. |
| Qin et al. (2010b) | Coupling of a core post-translational pacemaker to a slave transcription/translation feedback loop in a circadian system An ODE model that describes complex formation dynamics. The model neglects site-dependent phosphorylation and incorporates negative regulation of <i>kaiBC</i> transcription by KaiABC and KaiBC complexes. The model consists of 57 variables and 39 parameters. |
| Hertel et al. (2013) | Revealing a two-loop transcriptional feedback mechanism in the cyanobacterial circadian clock An extended version of the model of Brettschneider et al. (2010). The model includes positive and negative regulation of <i>kaiBC</i> transcription by threonine and doubly phosphorylated KaiC and unphosphorylated KaiC, respectively (15 variables, 33 parameters). |
| Teng et al. (2013) | Robust circadian oscillations in growing cyanobacteria require transcriptional feedback An extended version of the model of Rust et al. (2007). The model incorporates negative regulation of <i>kaiBC</i> transcription by serine phosphorylated KaiC (5 variables, 18 parameters). |

Later models combine the transcription–translation feedback loop (TTFL) and the core PTO oscillator. As for *in vitro* models, ATP, NADPH, and redox state are typically not considered as explicit variables. All models are implemented as ODEs, except otherwise noted.

A) have been identified, which reset or modulate the phase of KaiC phosphorylation cycle [reviewed by Mackey et al. (2011)]. **Figure 6** provides a schematic of possible sites of interactions. Mathematical models describing the interconnections at the molecular level have yet to be developed.

The cyanobacterial circadian clock results in genome-wide gene expression rhythms and regulates cell cycle progression relaying information via a two-component system that comprised SasA (*Synechococcus* adaptive sensor A) and RpaA (regulator of phycobilisome association A) (Liu et al., 1995; Mori et al., 1996; Takai et al., 2006; Dong et al., 2010). Rhythms of chromosome compaction and DNA topology (highly correlated with gene expression rhythms) do not hinge on SasA, pointing to the existence of other output pathways (Smith and Williams, 2006; Woelfle et al., 2007; Vijayan et al., 2009). In the positive transcriptional pathway, SasA interacts physically with KaiC and acts as a kinase toward RpaA (Takai et al., 2006; Gutu and OShea, 2013). In the course of a circadian cycle, KaiB displaces SasA from KaiC and KaiA becomes sequestered, switching KaiC into autodephosphorylation mode (**Figure 6**). The negative transcriptional output involves LabA (low amplitude and bright), CikA, and the transcriptional factor RpaB (regulator of phycobilisome associated B)—all three repressing the activity of RpaA (Taniguchi et al., 2007; Gutu and OShea, 2013; Espinosa et al., 2015). CikA, with its dual role in input and output pathways, plays a special role. As an output component, CikA competes with KaiA for binding to KaiB (Gutu and OShea, 2013; Chang et al., 2015). The binding of CikA to the KaiBC complex activates the phosphatase activity of CikA toward RpaA (Gutu and OShea, 2013) (**Figure 6**). RpaA, as both a circadian transcriptional activator and a repressor, drives global gene expression rhythms. The transcriptional output includes the regulation of clock genes, forming a transcription–translation feedback loop (TTFL) to the core oscillator (Ishiura et al., 1998). Since we are just beginning to understand the mechanistic details of the TTFL, the existing mathematical models are still highly simplified, using phenomenological assumptions as to how RpaA

(Zwicker et al., 2010) or specific phospho-forms of KaiC control transcription of the *kaiBC* gene cluster (Miyoshi et al., 2007; Qin et al., 2010b; Hertel et al., 2013; Teng et al., 2013). These models reproduce the most important experimental results, although Miyoshi et al. (2007) assumed hypothetical states for KaiA, KaiB, and KaiC inconsistent with experiments. Due the relatively small numbers of variables and parameters, the ODE models of Teng et al. (2013) and Hertel et al. (2013) might be most suitable for use in larger models of phototrophic growth but require additional modifications that account for the most recent experimental findings.

8. REGULATION OF GENE EXPRESSION IN CYANOBACTERIA

In *Synechococcus elongatus* PCC 7942, the KaiC phosphorylation cycle targets the general transcription apparatus and thereby regulates 30–100% of the transcriptome in circadian fashion, depending on the experimental setup (Liu et al., 1995; Nakahira et al., 2004; Ito et al., 2009; Vijayan et al., 2009; Lehmann et al., 2013). The transcriptional output is regulated by multiple factors such as circadian changes in chromosomal compaction/decompaction (Smith and Williams, 2006) involving oscillations in DNA supercoiling (Woelfle et al., 2007) as well as biochemical cascade pathways that converge to globally acting transcription factors, RpaA and RpaB and so far unknown factors (Taniguchi et al., 2007; Gutu and OShea, 2013; Paddock et al., 2013; Espinosa et al., 2015). Furthermore, it is now clear that small non-protein-coding RNAs (<200 nucleotides) play as both positive and negative regulators crucial roles in gene expression of cyanobacteria (Georg and Hess, 2011). By base-pairing with the target mRNA, small RNA molecules interfere with the ribosome binding site or other sequence stretches, and consequently alter mRNA translation and stability. This mode of regulation might explain why the proportion of cyclic proteins in diverse cyanobacteria is rather

uncorrelated to that found in microarray studies (Stöckel et al., 2011; Waldbauer et al., 2012; Guerreiro et al., 2014), because not only transcriptional but also post-transcriptional (small RNA-mediated) mechanisms might be active and modulate or fine-tune the dynamics of regulatory networks. *Synechocystis* sp. PCC 6803 possesses a large number of small non-coding RNAs, and antisense RNAs influence at least 26% of all gene transcripts in this cyanobacterium. There are several hints that the non-coding RNAs fulfill important functions in light–dark acclimation (Georg et al., 2009; Mitschke et al., 2011). A specific functional role could be clarified for some of the antisense RNAs, e.g., *IsrR* (Dühning et al., 2006; Legewie et al., 2008), *as-flv4* (Eisenhut et al., 2012), or *PsbAR2* and *PsbAR3* mRNA (Sakurai et al., 2012). Yet, many identified RNA regulators still await elucidation of their functional relevance.

9. CONCLUSION: PUTTING THE PARTS TOGETHER

Almost all cellular functions have evolved in the context of constraints and trade-offs that can only be understood if the respective cellular and environmental context is taken into account. To this end, the construction of computational models of cellular processes not only allows us to study the inner workings of selected processes but also allows us to investigate the emergent properties that arise from interactions between these processes. The trade-offs and interrelations within phototrophic growth are manifold: the energy required for cellular growth is derived from the photosynthetic light reactions, which themselves are a major source of reactive oxygen species (ROS) and therefore require careful balance between different electron transport pathways and alternative electron “valves”. CCMs use energy and have implications for the efficiency of carbon assimilation. The cellular ATP/ADP ratio and the oxidized PQ pool relays information to the circadian clock, which affects transcriptional output and hence metabolic activity. Metabolism itself depends on cellular energy and redox potential—and must be appropriately coordinated to synthesize the right metabolites at the right time. The availability of ribosomes and amino acids, which are itself products of metabolism, affects the rates of translation of new proteins—which must be coordinated to account for damage, stability, and turnover times of proteins. In particular, the components of PSII complexes themselves are dependent on environmental conditions due to photodamage caused by ROS (Yao et al., 2012).

Many if not most of these interactions and trade-offs are still insufficiently understood. An important example is the action and the evolutionary benefit of the circadian clock. While many if not most prokaryotes live in environments with periodic diurnal cycles of light, temperature, and humidity, only cyanobacteria are known to possess a *bona fide* circadian clock. While the competitive advantage of a circadian clock in a periodic environment has been demonstrated experimentally for cyanobacteria (Woelfle et al., 2004), the precise adaptive value and the selective pressure resulting the evolution of a clock remains only partially understood. Reasoning about the possible fitness implications of a circadian clock necessarily involves considering the organisms as a whole, as exemplified in the “escape from light” hypothesis

that circadian rhythmicity arose from the need to protect the organism’s DNA from ultraviolet (UV) radiation, at the time unfiltered by the Earth’s early atmosphere (Hut and Beersma, 2011; Lück and Westermarck, 2016). A quantitative evaluation of such a hypothesis requires to contrast the energetic cost of the circadian clock with its benefits for survival and growth—a task where advanced computational models will allow for an increasingly quantitative evaluation.

While, as outlined in this contribution, a large number of computational models related to phototrophic growth are already available, also many important cellular processes are still insufficiently described. An important example is the coordination of cellular growth in response to transient darkness, starvation, or stress conditions. Only recently, iconic pathways, such as the stringent response, have been shown to be active also in cyanobacteria and to mediate a coordinated transcriptional and translational reaction to (transient) periods of darkness (Hood et al., 2016). Likewise, knowledge about the molecular and physiological mechanisms involved in the transition of cyanobacterial cells from a resting state to an active vegetative state is still incomplete (Klotz et al., 2016)—albeit crucial to understand what mechanisms causes a cyanobacterial cell to divert resources away from growth and division and toward survival until environmental conditions improve. Also, only little is known concerning the coordination of cellular metabolism with cell cycle events (Asato, 2005, 2006). In particular, cell size control and size homeostasis in bacteria is still not fully understood, with conceptual models dating back to the work of Donachie (1968). Recent work on *E. coli* and *B. subtilis* favored the “adder” model, in which the size added between birth and division is constant for a given growth condition—as opposed to the “sizer,” in which the cell actively monitors cell size, and “timer” model, in which the cell grows for a specific time before division (Taheri-Araghi et al., 2015). Corresponding studies for cyanobacteria are not yet available. Recent data have indicated that there is coupling between circadian oscillator and the cell cycle, specifically that cell cycle progression in some cyanobacteria slows during specific circadian intervals (Dong et al., 2010; Yang et al., 2010)—posing timely questions for further computational research and necessitating integrative modules of cyanobacterial growth.

Overall, there is increasing interest in whole-cell models to understand cellular trade-offs and functions in the context of a living cells. The construction of integrative computational models to predict phenotype from genotype has gained momentum with a first whole-cell model of the life cycle of the human pathogen *Mycoplasma genitalium*—based on a subdivision of cell functionality into modules (Karr et al., 2012). A similar effort was already undertaken for cyanobacteria to explain fitness advantage conveyed by a circadian clock (Hellweger, 2010)—an approach that can be regarded as a prototype for the path outlined in this contribution.

Distinct from other approaches to whole-cell models, however, we argue that it is unlikely that a single universal model—a model that spans all scales from intracellular to intercellular to properties of ecosystems—will fulfill all requirements needed to describe cellular growth. Rather, we envision a modular approach. Depending on the research question, different temporal and

spatial scales must be considered. The task is then to derive an appropriate representation of cellular processes that accounts for the spatial and temporal scales involved. The derived submodels should be consistent with more fine-grained representation and their construction must be informed by knowledge how processes interact and which trade-offs are relevant for a specific research question. We note that, besides the biological challenges, such a strategy also entails major computational challenges. As yet, the annotation of computational models is often poor. That is, the biochemical identity of model variables is not defined in a computer-readable format, and hence, merging of models typically requires extensive manual curation (Krause et al., 2010). While standardized exchange formats for computational models, such as the Systems Biology Markup Language, SBML (Hucka et al., 2003), are available for more than a decade, they are not commonly applied outside the Systems Biology community. For example, as yet, none of the models of the cyanobacterial circadian clock are available from the BioModels database—a major resource for computational models of biological processes (Li et al., 2010). In addition, only few computational tools allow for the integration of different modeling concepts, such as constraint-based and kinetic models.

Notwithstanding these challenges in computational methodology, we expect a growing library of models related to cyanobacterial growth that can inform and guide research how phototrophic organisms, such as cyanobacteria, adapt to complex environments. These models must increasingly adhere to common standards and should be made available on open platforms, such as the BioModels database (Li et al., 2010) and e-cyanobacterium.org (Klement et al., 2013). Given the recent

advances in model development and annotation, computational modeling will undoubtedly play a key role in understanding trade-offs and adaptations in cyanobacteria. In the beginning, integrated models of cyanobacterial growth will be still idealistic, crude, and most certainly incomplete. But, again referring to Neidhardt (1999), “it is only through such modeling of whole-system behavior—that is, of growth—that one will learn how near and how far our knowledge takes us toward understanding the living cell.”

AUTHOR CONTRIBUTIONS

SW and RS conducted the research and wrote the article. Both authors approved the final manuscript.

ACKNOWLEDGMENTS

We would like to thank Jiří Jablonský for discussing his work on modeling photosynthetic light reactions and carbon metabolism, as well as Robert L. Burnap for comments and sharing insight about cyanobacterial physiology and growth.

FUNDING

This work is a contribution from the research grant *CyanoGrowth* funded by the German Federal Ministry of Education and Research as part of the e:Bio Innovationswettbewerb Systembiologie [e:Bio systems biology innovation competition] initiative (reference: FKZ 0316192).

REFERENCES

- Abe, J., Hiyama, T. B., Mukaiyama, A., Son, S., Mori, T., Saito, S., et al. (2015). Atomic-scale origins of slowness in the cyanobacterial circadian clock. *Science* 349, 312–316. doi:10.1126/science.1261040
- Acuña, A. M., Snellenburg, J. J., Gwizdala, M., Kirilovsky, D., van Grondelle, R., and van Stokkum, I. H. M. (2016). Resolving the contribution of the uncoupled phycoobilisomes to cyanobacterial pulse-amplitude modulated (PAM) fluorometry signals. *Photosynth. Res.* 127, 91–102. doi:10.1007/s11120-015-0141-x
- Allen, J. I., and Polimene, L. (2011). Linking physiology to ecology: towards a new generation of plankton models. *J. Plankton Res.* 33, 989–997. doi:10.1093/plankt/fbr032
- Amariei, C., Machné, R., Sasidharan, K., Gottstein, W., Tomita, M., Soga, T., et al. (2013). The dynamics of cellular energetics during continuous yeast culture. *Conf. Proc. IEEE Eng. Med. Biol. Soc.* 2013, 2708–2711. doi:10.1109/EMBC.2013.6610099
- Amarnath, K., Bennett, D. I. G., Schneider, A. R., and Fleming, G. R. (2016). Multiscale model of light harvesting by photosystem II in plants. *Proc. Natl. Acad. Sci. U.S.A.* 113, 1156–1161. doi:10.1073/pnas.1524999113
- Antoniewicz, M. R. (2013). Dynamic metabolic flux analysis—tools for probing transient states of metabolic networks. *Curr. Opin. Biotechnol.* 24, 973–978. doi:10.1016/j.copbio.2013.03.018
- Asato, Y. (2005). Control of ribosome synthesis during the cell division cycles of *E. coli* and *Synechococcus*. *Curr. Issues Mol. Biol.* 7, 109–117.
- Asato, Y. (2006). Perspectives in the coordinate regulation of cell cycle events in *Synechococcus*. *Curr. Issues Mol. Biol.* 8, 91–95.
- Axmann, I. M., Legewie, S., and Herzel, H. (2007). “A minimal circadian clock model,” in *Genome Inform.*, Vol. 18 (World Scientific), 54–64.
- Ayata, S.-D., Levy, M., Aumont, O., Sciandra, A., Sainte-Marie, J., Tagliabue, A., et al. (2013). Phytoplankton growth formulation in marine ecosystem models: should we take into account photo-acclimation and variable stoichiometry in oligotrophic areas? *J. Mar. Syst.* 125, 29–40. doi:10.1016/j.jmarsys.2012.12.010
- Badger, M. R., Bassett, M., and Comins, H. N. (1985). A model for HCO₃⁻ accumulation and photosynthesis in the cyanobacterium *Synechococcus* sp: theoretical predictions and experimental observations. *Plant Physiol.* 77, 465–471. doi:10.1104/pp.77.2.465
- Baroukh, C., Muoz-Tamayo, R., Steyer, J.-P., and Bernard, O. (2014). DRUM: a new framework for metabolic modeling under non-balanced growth. Application to the carbon metabolism of unicellular microalgae. *PLoS ONE* 9:e104499. doi:10.1371/journal.pone.0104499
- Beck, C., Hertel, S., Rediger, A., Lehmann, R., Wiegand, A., Kölsch, A., et al. (2014). Daily expression pattern of protein-encoding genes and small noncoding RNAs in *Synechocystis* sp. strain PCC 6803. *Appl. Environ. Microbiol.* 80, 5195–5206. doi:10.1128/AEM.01086-14
- Bernard, O. (2011). Hurdles and challenges challenges for modelling and control of microalgae for CO₂ mitigation and biofuel production. *J. Process Control* 21, 1378–1389. doi:10.1016/j.jprocont.2011.07.012
- Berry, S., and Rumberg, B. (2000). Kinetic modeling of the photosynthetic electron transport chain. *Bioelectrochemistry* 53, 35–53. doi:10.1016/S0302-4598(00)00108-2
- Brettschneider, C., Rose, R. J., Hertel, S., Axmann, I. M., Heck, A. J., and Kollmann, M. (2010). A sequestration feedback determines dynamics and temperature entrainment of the KaiABC circadian clock. *Mol. Syst. Biol.* 6, 389. doi:10.1038/msb.2010.44
- Burnap, R. L. (2015). Systems and photosystems: cellular limits of autotrophic productivity in cyanobacteria. *Front. Bioeng. Biotechnol.* 3:1. doi:10.3389/fbioe.2015.00001
- Burnap, R. L., Hagemann, M., and Kaplan, A. (2015). Regulation of CO₂ concentrating mechanism in cyanobacteria. *Life (Basel)* 5, 348–371. doi:10.3390/life5010348

- Butler, W. L. (1978). Energy distribution in the photochemical apparatus of photosynthesis. *Ann. Rev. Plant Physiol.* 29, 345–378. doi:10.1146/annurev.pp.29.060178.002021
- Butler, W. L., and Strasser, R. J. (1977). Tripartite model for the photochemical apparatus of green plant photosynthesis. *Proc. Natl. Acad. Sci. U.S.A.* 74, 3382–3385. doi:10.1073/pnas.74.8.3382
- Chakrabarti, A., Miskovic, L., Soh, K. C., and Hatzimanikatis, V. (2013). Towards kinetic modeling of genome-scale metabolic networks without sacrificing stoichiometric, thermodynamic and physiological constraints. *Biotechnol. J.* 8, 1043–1057. doi:10.1002/biot.201300091
- Chang, Y.-G., Cohen, S. E., Phong, C., Myers, W. K., Kim, Y.-I., Tseng, R., et al. (2015). A protein fold switch joins the circadian oscillator to clock output in cyanobacteria. *Science* 349, 324–328. doi:10.1126/science.1260031
- Cheung, C. Y. M., Poolman, M. G., Fell, D. A., Ratcliffe, R. G., and Sweetlove, L. J. (2014). A diel flux balance model captures interactions between light and dark metabolism during day-night cycles in C3 and crassulacean acid metabolism leaves. *Plant Physiol.* 165, 917–929. doi:10.1104/pp.113.234468
- Clark, R. L., Cameron, J. C., Root, T. W., and Pfleger, B. F. (2014). Insights into the industrial growth of cyanobacteria from a model of the carbon-concentrating mechanism. *AIChE J.* 60, 12691277. doi:10.1002/aic.14310
- Clodong, S., Dühring, U., Kronk, L., Wilde, A., Axmann, I., Herzel, H., et al. (2007). Functioning and robustness of a bacterial circadian clock. *Mol. Syst. Biol.* 3, 90. doi:10.1038/msb4100128
- Crill, P. A. (1977). The photosynthesis-light curve: a simple analog model. *J. Theor. Biol.* 64, 503–516. doi:10.1016/0022-5193(77)90284-3
- Diamond, S., Jun, D., Rubin, B. E., and Golden, S. S. (2015). The circadian oscillator in *Synechococcus elongatus* controls metabolite partitioning during diurnal growth. *Proc. Natl. Acad. Sci. U.S.A.* 112, E1916–E1925. doi:10.1073/pnas.1504576112
- Donachie, W. D. (1968). Relationship between cell size and time of initiation of DNA replication. *Nature* 219, 1077–1079. doi:10.1038/2191077a0
- Dong, G., Yang, Q., Wang, Q., Kim, Y.-I., Wood, T. L., Osteryoung, K. W., et al. (2010). Elevated ATPase activity of KaiC regulates a circadian checkpoint on cell division in *Synechococcus elongatus*. *Cell* 140, 529–539. doi:10.1016/j.cell.2009.12.042
- Droop, M. R. (1968). Vitamin B12 and marine ecology. IV. The kinetics of uptake, growth and inhibition in *Monochrysis lutheri*. *J. Mar. Biol. Assoc. U.K.* 48, 689–733. doi:10.1017/S0025315400019238
- Droop, M. R. (1983). 25 years of algal growth kinetic: a personal view. *Botanica Marina* XXVI, 99–112.
- Dühring, U., Axmann, I. M., Hess, W. R., and Wilde, A. (2006). An internal antisense RNA regulates expression of the photosynthesis gene *isiA*. *Proc. Natl. Acad. Sci. U.S.A.* 103, 7054–7058. doi:10.1073/pnas.0600927103
- Ebenhöh, O., Houwaart, T., Lokstein, H., Schleder, S., and Tirok, K. (2011). A minimal mathematical model of nonphotochemical quenching of chlorophyll fluorescence. *BioSystems* 103, 196–204. doi:10.1016/j.biosystems.2010.10.011
- Egli, M., Mori, T., Pattanayek, R., Xu, Y., Qin, X., and Johnson, C. H. (2012). Dephosphorylation of the core clock protein KaiC in the cyanobacterial KaiABC circadian oscillator proceeds via an ATP synthase mechanism. *Biochemistry* 51, 1547–1558. doi:10.1021/bi201525n
- Eguchi, K., Yoda, M., Terada, T. P., and Sasai, M. (2008). Mechanism of robust circadian oscillation of KaiC phosphorylation in vitro. *Biophys. J.* 95, 1773–1784. doi:10.1529/biophysj.107.127555
- Eilers, P. H. C., and Peeters, J. C. H. (1988). A model for the relationship between light intensity and the rate of photosynthesis. *Ecol. Model.* 42, 199–215. doi:10.1016/0304-3800(88)90057-9
- Eisenhut, M., Georg, J., Klähn, S., Sakurai, I., Mustila, H., Zhang, P., et al. (2012). The antisense RNA *As1_flv4* in the cyanobacterium *Synechocystis* sp. PCC 6803 prevents premature expression of the *flv4-2* operon upon shift in inorganic carbon supply. *J. Biol. Chem.* 287, 33153–33162. doi:10.1074/jbc.M112.391755
- Emberly, E., and Wingreen, N. S. (2006). Hourglass model for a protein-based circadian oscillator. *Phys. Rev. Lett.* 96, 038303. doi:10.1103/PhysRevLett.96.038303
- Erdreich, P., Steuer, R., and Klamt, S. (2015). An algorithm for the reduction of genome-scale metabolic network models to meaningful core models. *BMC Syst. Biol.* 9:48. doi:10.1186/s12918-015-0191-x
- Espinosa, J., Boyd, J. S., Cantos, R., Salinas, P., Golden, S. S., and Contreras, A. (2015). Cross-talk and regulatory interactions between the essential response regulator RpaB and cyanobacterial circadian clock output. *Proc. Natl. Acad. Sci. U.S.A.* 112, 2198–2203. doi:10.1073/pnas.1424632112
- Farquhar, G. D., von Caemmerer, S., and Berry, J. A. (1980). A biochemical model of photosynthetic CO₂ assimilation in leaves of C3 species. *Planta* 149, 78–90. doi:10.1007/BF00386231
- Feng, X., Xu, Y., Chen, Y., and Tang, Y. J. (2012). Integrating flux balance analysis into kinetic models to decipher the dynamic metabolism of *Shewanella oneidensis* mr-1. *PLoS Comput. Biol.* 8:e1002376. doi:10.1371/journal.pcbi.1002376
- Fokina, O., Chellamuthu, V.-R., Forchhammer, K., and Zeth, K. (2010). Mechanism of 2-oxoglutarate signaling by the *Synechococcus elongatus* PII signal transduction protein. *Proc. Natl. Acad. Sci. U.S.A.* 107, 19760–19765. doi:10.1073/pnas.1007653107
- Fridlyand, L., Kaplan, A., and Reinhold, L. (1996). Quantitative evaluation of the role of a putative CO₂-scavenging entity in the cyanobacterial CO₂-concentrating mechanism. *BioSystems* 37, 229–238. doi:10.1016/0303-2647(95)01561-2
- Georg, J., and Hess, W. R. (2011). cis-Antisense RNA, another level of gene regulation in bacteria. *Microbiol. Mol. Biol. Rev.* 75, 286–300. doi:10.1128/MMBR.00032-10
- Georg, J., Voß, B., Scholz, I., Mitschke, J., Wilde, A., and Hess, W. R. (2009). Evidence for a major role of antisense RNAs in cyanobacterial gene regulation. *Mol. Syst. Biol.* 5, 305. doi:10.1038/msb.2009.63
- Giersch, C. (1986). Oscillatory response of photosynthesis in leaves to environmental perturbations: a mathematical model. *Arch. Biochem. Biophys.* 245, 263–270. doi:10.1016/0003-9861(86)90213-4
- Goel, A., Wortel, M. T., Molenaar, D., and Teusink, B. (2012). Metabolic shifts: a fitness perspective for microbial cell factories. *Biotechnol. Lett.* 34, 2147–2160. doi:10.1007/s10529-012-1038-9
- Gorbunov, M. Y., Kuzminov, F. I., Fadeev, V. V., Kim, J. D., and Falkowski, P. G. (2011). A kinetic model of non-photochemical quenching in cyanobacteria. *Biochim. Biophys. Acta* 1807, 1591–1599. doi:10.1016/j.bbabi.2011.08.009
- Guerreiro, A. C., Benevento, M., Lehmann, R., van Breukelen, B., Post, H., Giansanti, P., et al. (2014). Daily rhythms in the cyanobacterium *Synechococcus elongatus* probed by high-resolution mass spectrometry-based proteomics reveals a small defined set of cyclic proteins. *Mol. Cell. Proteomics* 13, 2042–2055. doi:10.1074/mcp.M113.035840
- Gutu, A., and O'Shea, E. K. (2013). Two antagonistic clock-regulated histidine kinases time the activation of circadian gene expression. *Mol. Cell* 50, 288–294. doi:10.1016/j.molcel.2013.02.022
- Hahn, B. D. (1987). A mathematical model of photorespiration and photosynthesis. *Ann. Bot.* 60, 157–169.
- Hahn, B. D. (1991). Photosynthesis and photorespiration: modelling the essentials. *J. Theor. Biol.* 151, 123–139. doi:10.1016/S0022-5193(05)80147-X
- Hahn, D. B. (1984). A mathematical model of leaf carbon metabolism. *Ann. Bot.* 54, 325–339.
- Hamilton, J. J., and Reed, J. L. (2012). Identification of functional differences in metabolic networks using comparative genomics and constraint-based models. *PLoS ONE* 7:e34670. doi:10.1371/journal.pone.0034670
- Han, B. P. (2001). Photosynthesis-irradiance response at physiological level: a mechanistic model. *J. Theor. Biol.* 213, 121–127. doi:10.1006/jtbi.2001.2413
- Han, B.-P. (2002). A mechanistic model of algal photoinhibition induced by photodamage to photosystem-II. *J. Theor. Biol.* 214, 519–527. doi:10.1006/jtbi.2001.2468
- Hellweger, F. L. (2010). Resonating circadian clock enhance fitness in cyanobacteria in silico. *Ecol. Model.* 221, 1620–1629. doi:10.1016/j.ecolmodel.2010.03.015
- Henson, M. A. (2015). Genome-scale modelling of microbial metabolism with temporal and spatial resolution. *Biochem. Soc. Trans.* 43, 1164–1171. doi:10.1042/BST20150146
- Hertel, S., Brettschneider, C., and Axmann, I. M. (2013). Revealing a two-loop transcriptional feedback mechanism in the cyanobacterial circadian clock. *PLoS Comput. Biol.* 9:e1002966. doi:10.1371/journal.pcbi.1002966
- Holzwarth, A. R., Müller, M. G., Reus, M., Nowaczyk, M., Sander, J., and Rögner, M. (2006). Kinetics and mechanism of electron transfer in intact photosystem II

- and in the isolated reaction center: pheophytin is the primary electron acceptor. *Proc. Natl. Acad. Sci. U.S.A.* 103, 6895–6900. doi:10.1073/pnas.0505371103
- Hood, R. D., Higgins, S. A., Flamholz, A., Nichols, R. J., and Savage, D. F. (2016). The stringent response regulates adaptation to darkness in the cyanobacterium *Synechococcus elongatus*. *Proc. Natl. Acad. Sci. U.S.A.* 113, E4867–E4876. doi:10.1073/pnas.1524915113
- Hopkinson, B. M., Young, J. N., Tansik, A. L., and Binder, B. J. (2014). The minimal CO₂-concentrating mechanism of *Prochlorococcus* spp. MED4 is effective and efficient. *Plant Physiol.* 166, 2205–2217. doi:10.1104/pp.114.247049
- Hucka, M., Finney, A., Sauro, H. M., Bolouri, H., Doyle, J. C., Kitano, H., et al. (2003). The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models. *Bioinformatics* 19, 524–531. doi:10.1093/bioinformatics/btg015
- Hut, R. A., and Beersma, D. G. M. (2011). Evolution of time-keeping mechanisms: early emergence and adaptation to photoperiod. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 366, 2141–2154. doi:10.1098/rstb.2010.0409
- Ishiura, M., Kutsuna, S., Aoki, S., Iwasaki, H., Andersson, C. R., Tanabe, A., et al. (1998). Expression of a gene cluster kaiABC as a circadian feedback process in cyanobacteria. *Science* 281, 1519–1523. doi:10.1126/science.281.5382.1519
- Ito, H., Mutsuda, M., Murayama, Y., Tomita, J., Hosokawa, N., Terauchi, K., et al. (2009). Cyanobacterial daily life with Kai-based circadian and diurnal genome-wide transcriptional control in *Synechococcus elongatus*. *Proc. Natl. Acad. Sci. U.S.A.* 106, 14168–14173. doi:10.1073/pnas.0902587106
- Iwasaki, H., Nishiwaki, T., Kitayama, Y., Nakajima, M., and Kondo, T. (2002). KaiA-stimulated KaiC phosphorylation in circadian timing loops in cyanobacteria. *Proc. Natl. Acad. Sci. U.S.A.* 99, 15788–15793. doi:10.1073/pnas.222467299
- Jablonský, J., Hagemann, M., Schwarz, D., and Wolkenhauer, O. (2013). Phosphoglycerate mutases function as reverse regulated isoenzymes in *Synechococcus elongatus* PCC 7942. *PLoS ONE* 8:e58281. doi:10.1371/journal.pone.0058281
- Jablonský, J., and Lázár, D. (2008). Evidence for intermediate S-states as initial phase in the process of oxygen-evolving complex oxidation. *Biophys. J.* 94, 2725–2736. doi:10.1529/biophysj.107.122861
- Jablonský, J., Papacek, S., and Hagemann, M. (2016). Different strategies of metabolic regulation in cyanobacteria: from transcriptional to biochemical control. *Sci. Rep.* 6, 33024. doi:10.1038/srep33024
- Jablonský, J., Schwarz, D., and Hagemann, M. (2014). Multi-level kinetic model explaining diverse roles of isozymes in prokaryotes. *PLoS ONE* 9:e105292. doi:10.1371/journal.pone.0105292
- Jablonský, J., Susila, P., and Lázár, D. (2008). Impact of dimeric organization of enzyme on its function: the case of photosynthetic water splitting. *Bioinformatics* 24, 2755–2759. doi:10.1093/bioinformatics/btn530
- Kageyama, H., Nishiwaki, T., Nakajima, M., Iwasaki, H., Oyama, T., and Kondo, T. (2006). Cyanobacterial circadian pacemaker: Kai protein complex dynamics in the KaiC phosphorylation cycle in vitro. *Mol. Cell* 23, 161–171. doi:10.1016/j.molcel.2006.05.039
- Kaplan, A., and Reinhold, L. (1999). CO₂ concentrating mechanisms in photosynthetic microorganisms. *Annu. Rev. Plant Physiol. Plant Mol. Biol.* 50, 539–570. doi:10.1146/annurev.arplant.50.1.539
- Karr, J. R., Sanghvi, J. C., Macklin, D. N., Gutschow, M. V., Jacobs, J. M., Bolival, B. Jr., et al. (2012). A whole-cell computational model predicts phenotype from genotype. *Cell* 150, 389–401. doi:10.1016/j.cell.2012.05.044
- Kim, J. I., Varner, J. D., and Ramkrishna, D. (2008). A hybrid model of anaerobic *E. coli* GJT001: combination of elementary flux modes and cybernetic variables. *Biotechnol. Prog.* 24, 993–1006. doi:10.1002/btpr.73
- Kim, Y.-I., Vinyard, D. J., Ananyev, G. M., Dismukes, G. C., and Golden, S. S. (2012). Oxidized quinones signal onset of darkness directly to the cyanobacterial circadian oscillator. *Proc. Natl. Acad. Sci. U.S.A.* 109, 17765–17769. doi:10.1073/pnas.1216401109
- Kirilovsky, D., and Kerfeld, C. A. (2013). The orange carotenoid protein: a blue-green light photoactive protein. *Photochem. Photobiol. Sci.* 12, 1135–1143. doi:10.1039/c3pp25406b
- Kitayama, Y., Iwasaki, H., Nishiwaki, T., and Kondo, T. (2003). KaiB functions as an attenuator of KaiC phosphorylation in the cyanobacterial circadian clock system. *EMBO J.* 22, 2127–2134. doi:10.1093/emboj/cdg212
- Klement, M., Safranek, D., Ded, T., Pejznoch, A., Nedbal, L., Steuer, R., et al. (2013). A comprehensive web-based platform for domain-specific biological models. *ENTCS* 299, 61–67. doi:10.1016/j.entcs.2013.11.006
- Klotz, A., Georg, J., Bučinská, L., Watanabe, S., Reimann, V., Januszewski, W., et al. (2016). Awakening of a dormant cyanobacterium from nitrogen chlorosis reveals a genetically determined program. *Curr. Biol.* 26, 2862–2872. doi:10.1016/j.cub.2016.08.054
- Knies, D., Wittmüß, P., Appel, S., Sawodny, O., Ederer, M., and Feuer, R. (2015). Modeling and simulation of optimal resource management during the diurnal cycle in *Emiliania huxleyi* by genome-scale reconstruction and an extended flux balance analysis approach. *Metabolites* 5, 659–676. doi:10.3390/metabo5040659
- Knoop, H., Gründel, M., Zilliges, Y., Lehmann, R., Hoffmann, S., Lockau, W., et al. (2013). Flux balance analysis of cyanobacterial metabolism: the metabolic network of *Synechocystis* sp. PCC 6803. *PLoS Comput. Biol.* 9:e1003081. doi:10.1371/journal.pcbi.1003081
- Knoop, H., Zilliges, Y., Lockau, W., and Steuer, R. (2010). The metabolic network of *Synechocystis* sp. PCC 6803: systemic properties of autotrophic growth. *Plant Physiol.* 154, 410–422. doi:10.1104/pp.110.157198
- Kok, B., Forbush, B., and McGloin, M. (1970). Cooperation of charges in photosynthetic O₂ evolution-I. A linear four step mechanism. *Photochem. Photobiol.* 11, 457475. doi:10.1111/j.1751-1097.1970.tb06017.x
- Kopf, M., and Hess, W. R. (2015). Regulatory RNAs in photosynthetic cyanobacteria. *FEMS Microbiol. Rev.* 39, 301–315. doi:10.1093/femsre/fuv017
- Kopf, M., Klähn, S., Scholz, I., Matthiessen, J. K. F., Hess, W. R., and Voß, B. (2014). Comparative analysis of the primary transcriptome of *Synechocystis* sp. PCC 6803. *DNA Res.* 21, 527–539. doi:10.1093/dnares/dsu018
- Krause, F., Uhlenendorf, J., Lubitz, T., Schulz, M., Klipp, E., and Liebermeister, W. (2010). Annotation and merging of SBML models with semanticSBML. *Bioinformatics* 26, 421–422. doi:10.1093/bioinformatics/btp642
- Kroon, B. M. A., and Thoms, S. (2006). From electron to biomass: a mechanistic model to describe phytoplankton photosynthesis and steady-state growth rates. *J. Phycol.* 42, 593–609. doi:10.1111/j.1529-8817.2006.00221.x
- Kurosawa, G., Aihara, K., and Iwasa, Y. (2006). A model for the circadian rhythm of cyanobacteria that maintains oscillation without gene expression. *Biophys. J.* 91, 2015–2023. doi:10.1529/biophysj.105.076554
- Laisk, A., Eichelmann, H., and Oja, V. (2006). C3 photosynthesis in silico. *Photosynth. Res.* 90, 45–66. doi:10.1007/s11220-006-9109-1
- Laisk, A., Eichelmann, H., Oja, V., Eatherall, A., and Walker, D. A. (1989). A mathematical model of the carbon metabolism in photosynthesis. difficulties in explaining oscillations by fructose 2,6-bisphosphate regulation. *Proc. R. Soc. Lond. B* 237, 389–415. doi:10.1098/rspb.1989.0057
- Laisk, A., and Walker, D. A. (1986). Control of phosphate turnover as a rate-limiting factor and possible cause of oscillations in photosynthesis: a mathematical model. *Proc. R. Soc. Lond. B* 227, 281–302. doi:10.1098/rspb.1986.0024
- Lane, N. (2002). *Oxygen: The molecule That Made the World*. Oxford: Oxford University Press.
- Lázár, D. (2003). Chlorophyll a fluorescence rise induced by high light illumination of dark-adapted plant tissue studied by means of a model of photosystem II and considering photosystem II heterogeneity. *J. Theor. Biol.* 220, 469–503. doi:10.1006/jtbi.2003.3140
- Lázár, D., and Jablonský, J. (2009). On the approaches applied in formulation of a kinetic model of photosystem II: different approaches lead to different simulations of the chlorophyll alpha fluorescence transients. *J. Theor. Biol.* 257, 260–269. doi:10.1016/j.jtbi.2008.11.018
- Legewie, S., Dienst, D., Wilde, A., Herzel, H., and Axmann, I. M. (2008). Small RNAs establish delays and temporal thresholds in gene expression. *Biophys. J.* 95, 3232–3238. doi:10.1529/biophysj.108.133819
- Lehmann, R., Machné, R., Georg, J., Benary, M., Axmann, I., and Steuer, R. (2013). How cyanobacteria pose new problems to old methods: challenges in microarray time series analysis. *BMC Bioinformatics* 14:133. doi:10.1186/1471-2105-14-133
- Li, C., Donizelli, M., Rodriguez, N., Dharuri, H., Endler, L., Chelliah, V., et al. (2010). Biomodels database: an enhanced, curated and annotated resource for published quantitative kinetic models. *BMC Syst. Biol.* 4:92. doi:10.1186/1752-0509-4-92
- Liu, Y., Tsinoremas, N. F., Johnson, C. H., Lebedeva, N. V., Golden, S. S., Ishiura, M., et al. (1995). Circadian orchestration of gene expression in cyanobacteria. *Genes Dev.* 9, 1469–1478. doi:10.1101/gad.9.12.1469
- Lück, S., and Westermarck, P. O. (2016). Circadian mRNA expression: insights from modeling and transcriptomics. *Cell Mol. Life Sci.* 73, 497–521. doi:10.1007/s00018-015-2072-2

- Ma, W., and Ogawa, T. (2015). Oxygenic photosynthesis-specific subunits of cyanobacterial NADPH dehydrogenases. *IUBMB Life* 67, 3–8. doi:10.1002/iub.1341
- Maarelveld, T. R., Boele, J., Bruggeman, F. J., and Teusink, B. (2014). A data integration and visualization resource for the metabolic network of *Synechocystis* sp. PCC 6803. *Plant Physiol.* 164, 1111–1121. doi:10.1104/pp.113.224394
- Mackey, S. R., Golden, S. S., and Ditty, J. L. (2011). The itty-bitty time machine: genetics of the cyanobacterial circadian clock. *Adv. Genet.* 74, 13. doi:10.1016/B978-0-12-387690-4.00002-7
- Mahadevan, R., Edwards, J. S., and Doyle, F. J. III (2002). Dynamic flux balance analysis of diauxic growth in *Escherichia coli*. *Biophys. J.* 83, 1331–1340. doi:10.1016/S0006-3495(02)73903-9
- Mangan, N., and Brenner, M. (2014). Systems analysis of the CO₂ concentrating mechanism in cyanobacteria. *eLife* 3:e02043. doi:10.7554/eLife.02043
- Mangan, N. M., Flamholz, A., Hood, R. D., Milo, R., and Savage, D. F. (2016). pH determines the energetic efficiency of the cyanobacterial CO₂ concentrating mechanism. *Proc. Natl. Acad. Sci. U.S.A.* 113, E5354–E5362. doi:10.1073/pnas.1525145113
- Mar, T., and Govindjee (1972). Kinetic models of oxygen evolution in photosynthesis. *J. Theor. Biol.* 36, 427–446. doi:10.1016/0022-5193(72)90001-X
- McCloskey, D., Palsson, B. O., and Feist, A. M. (2013). Basic and applied uses of genome-scale metabolic network reconstructions of *Escherichia coli*. *Mol. Syst. Biol.* 9, 661. doi:10.1038/msb.2013.18
- Megard, R. O., Tonkyn, D. W., and Senft, W. H. (1984). Kinetics of oxygenic photosynthesis in plankton algae. *J. Plankton Res.* 6, 325–337. doi:10.1093/plankt/6.2.325
- Mehra, A., Hong, C. I., Shi, M., Loros, J. J., Dunlap, J. C., and Ruoff, P. (2006). Circadian rhythmicity by autocatalysis. *PLoS Comput. Biol.* 2:e96. doi:10.1371/journal.pcbi.0020096
- Milstein, J., and Bremermann, H. J. (1979). Parameter identification of the Calvin photosynthetic cycle. *J. Math. Biol.* 7, 99–116. doi:10.1007/BF00276924
- Mitschke, J., Georg, J., Scholz, I., Sharma, C. M., Dienst, D., Bantscheff, J., et al. (2011). An experimentally anchored map of transcriptional start sites in the model cyanobacterium *Synechocystis* sp. PCC 6803. *Proc. Natl. Acad. Sci. U.S.A.* 108, 2124–2129. doi:10.1073/pnas.1015154108
- Miyoshi, F., Nakayama, Y., Kaizu, K., Iwasaki, H., and Tomita, M. (2007). A mathematical model for the Kai-protein-based chemical oscillator and clock gene expression rhythms in cyanobacteria. *J. Biol. Rhythms* 22, 69–80. doi:10.1177/0748730406295749
- Molenaar, D., van Berlo, R., de Ridder, D., and Teusink, B. (2009). Shifts in growth strategies reflect tradeoffs in cellular economics. *Mol. Syst. Biol.* 5, 323. doi:10.1038/msb.2009.82
- Montagud, A., Navarro, E., Fernandez de Cordoba, P., Urchueguia, J. F., and Patil, K. R. (2010). Reconstruction and analysis of genome-scale metabolic model of a photosynthetic bacterium. *BMC Syst. Biol.* 4:156. doi:10.1186/1752-0509-4-156
- Mori, T., Binder, B., and Johnson, C. H. (1996). Circadian gating of cell division in cyanobacteria growing with average doubling times of less than 24 hours. *Proc. Natl. Acad. Sci. U.S.A.* 93, 10183–10188. doi:10.1073/pnas.93.19.10183
- Mori, T., Williams, D. R., Byrne, M. O., Qin, X., Egli, M., Mchaurab, H. S., et al. (2007). Elucidating the ticking of an in vitro circadian clockwork. *PLoS Biol.* 5:e93. doi:10.1371/journal.pbio.0050093
- Morton, O. (2009). *Eating the Sun: How Plants Power the Planet*. New York: HarperCollins Publishers.
- Mueller, T. J., Berla, B. M., Pakrasi, H. B., and Maranas, C. D. (2013). Rapid construction of metabolic models for a family of cyanobacteria using a multiple source annotation workflow. *BMC Syst. Biol.* 7:142. doi:10.1186/1752-0509-7-142
- Müller, S., Regensburger, G., and Steuer, R. (2015). Resource allocation in metabolic networks: kinetic optimization and approximations by FBA. *Biochem. Soc. Trans.* 43, 1195–1200. doi:10.1042/BST20150156
- Mullineaux, C. W. (2014). Electron transport and light-harvesting switches in cyanobacteria. *Front. Plant Sci.* 5:7. doi:10.3389/fpls.2014.00007
- Murabito, E., Verma, M., Bekker, M., Bellomo, D., Westerhoff, H. V., Teusink, B., et al. (2014). Monte-Carlo modeling of the central carbon metabolism of *Lactococcus lactis*: insights into metabolic regulation. *PLoS ONE* 9:e106453. doi:10.1371/journal.pone.0106453
- Muthuraj, M., Palabhanvi, B., Misra, S., Kumar, V., Sivalingavas, K., and Das, D. (2013). Flux balance analysis of *Chlorella* sp. FC2 IITG under photoautotrophic and heterotrophic growth conditions. *Photosynth. Res.* 118, 167–179. doi:10.1007/s11120-013-9943-x
- Nagai, T., Terada, T. P., and Sasai, M. (2010). Synchronization of circadian oscillation of phosphorylation level of KaiC in vitro. *Biophys. J.* 98, 2469–2477. doi:10.1016/j.bpj.2010.02.036
- Nakahira, Y., Katayama, M., Miyashita, H., Kutsuna, S., Iwasaki, H., Oyama, T., et al. (2004). Global gene repression by KaiC as a master process of prokaryotic circadian system. *Proc. Natl. Acad. Sci. U.S.A.* 101, 881–885. doi:10.1073/pnas.0307411100
- Nakajima, M., Imai, K., Ito, H., Nishiwaki, T., Murayama, Y., Iwasaki, H., et al. (2005). Reconstitution of circadian oscillation of cyanobacterial KaiC phosphorylation in vitro. *Science* 308, 414–415. doi:10.1126/science.1108451
- Nedbal, L., Cerveny, J., Keren, N., and Kaplan, A. (2010). Experimental validation of a nonequilibrium model of CO₂ fluxes between gas, liquid medium, and algae in a flat-panel photobioreactor. *J. Ind. Microbiol. Biotechnol.* 37, 1319–1326. doi:10.1007/s10295-010-0876-5
- Nedbal, L., Cerveny, J., Rascher, U., and Schmidt, H. (2007). E-photosynthesis: a comprehensive modeling approach to understand chlorophyll fluorescence transients and other complex dynamic features of photosynthesis in fluctuating light. *Photosynth. Res.* 93, 223–234. doi:10.1007/s11120-007-9178-9
- Neidhardt, F. C. (1999). Bacterial growth: constant obsession with dN/dt. *J. Bacteriol.* 181, 7405–7408.
- Neidhardt, F. C., Ingraham, J. L., and Schaechter, M. (1990). *Physiology of the Bacterial Cell: A Molecular Approach*. Sunderland, MA: Sinauer Associates, Inc.
- Nishiwaki, T., and Kondo, T. (2012). Circadian autodephosphorylation of cyanobacterial clock protein KaiC occurs via formation of ATP as intermediate. *J. Biol. Chem.* 287, 18030–18035. doi:10.1074/jbc.M112.350660
- Nishiwaki, T., Satomi, Y., Kitayama, Y., Terauchi, K., Kiyohara, R., Takao, T., et al. (2007). A sequential program of dual phosphorylation of KaiC as a basis for circadian rhythm in cyanobacteria. *EMBO J.* 26, 4029–4037. doi:10.1038/sj.emboj.7601832
- Nishiwaki, T., Satomi, Y., Nakajima, M., Lee, C., Kiyohara, R., Kageyama, H., et al. (2004). Role of KaiC phosphorylation in the circadian clock system of *Synechococcus elongatus* PCC 7942. *Proc. Natl. Acad. Sci. U.S.A.* 101, 13927–13932. doi:10.1073/pnas.0403906101
- Nogales, J., Gudmundsson, S., Knight, E. M., Palsson, B. O., and Thiele, I. (2012). Detailing the optimality of photosynthesis in cyanobacteria through systems biology analysis. *Proc. Natl. Acad. Sci. U.S.A.* 109, 2678–2683. doi:10.1073/pnas.1117907109
- Paddock, M. L., Boyd, J. S., Adin, D. M., and Golden, S. S. (2013). Active output state of the *Synechococcus* Kai circadian oscillator. *Proc. Natl. Acad. Sci. U.S.A.* 110, E3849–E3857. doi:10.1073/pnas.1315170110
- Pattanaik, G., and Rust, M. J. (2014). The cyanobacterial clock and metabolism. *Curr. Opin. Microbiol.* 18, 90–95. doi:10.1016/j.mib.2014.02.010
- Pattanaik, G. K., Lambert, G., Bernat, K., and Rust, M. J. (2015). Controlling the cyanobacterial clock by synthetically rewiring metabolism. *Cell Rep.* 13, 2362–2367. doi:10.1016/j.celrep.2015.11.031
- Pattanaik, G. K., Phong, C., and Rust, M. J. (2014). Rhythms in energy storage control the ability of the cyanobacterial circadian clock to reset. *Curr. Biol.* 24, 1934–1938. doi:10.1016/j.cub.2014.07.022
- Petersson, G., and Ryde-Petersson, U. (1988). A mathematical model of the Calvin photosynthesis cycle. *Eur. J. Biochem.* 175, 661–672. doi:10.1111/j.1432-1033.1988.tb14242.x
- Phong, C., Markson, J. S., Wilhoite, C. M., and Rust, M. J. (2013). Robust and tunable circadian rhythms from differentially sensitive catalytic domains. *Proc. Natl. Acad. Sci. U.S.A.* 110, 1124–1129. doi:10.1073/pnas.1212113110
- Poolman, M. G., Fell, D. A., and Thomas, S. (2000). Modelling photosynthesis and its control. *J. Exp. Bot.* 51 Spec No, 319–328. doi:10.1093/jexbot/51.suppl_1.319
- Poolman, M. G., Olcer, H., Lloyd, J. C., Raines, C. A., and Fell, D. A. (2001). Computer modelling and experimental evidence for two steady states in the photosynthetic Calvin cycle. *Eur. J. Biochem.* 268, 2810–2816. doi:10.1046/j.1432-1327.2001.02153.x
- Price, G. D., Badger, M. R., Woodger, F. J., and Long, B. M. (2008). Advances in understanding the cyanobacterial CO₂-concentrating-mechanism (CCM): functional components, Ci transporters, diversity, genetic regulation and prospects for engineering into plants. *J. Exp. Bot.* 59, 1441–1461. doi:10.1093/jxb/erm112

- Qin, X., Byrne, M., Mori, T., Zou, P., Williams, D. R., Mchaourab, H., et al. (2010a). Intermolecular associations determine the dynamics of the circadian KaiABC oscillator. *Proc. Natl. Acad. Sci. U.S.A.* 107, 14805–14810. doi:10.1073/pnas.1002119107
- Qin, X., Byrne, M., Xu, Y., Mori, T., and Johnson, C. H. (2010b). Coupling of a core post-translational pacemaker to a slave transcription/translation feedback loop in a circadian system. *PLoS Biol.* 8:1310. doi:10.1371/journal.pbio.1000394
- Reinhold, L., Kosloff, R., and Kaplan, A. (1991). A model for inorganic carbon fluxes and photosynthesis in cyanobacterial carboxysomes. *Can. J. Bot.* 69, 984–988. doi:10.1139/b91-126
- Reinhold, L., Zviman, M., and Kaplan, A. (1987). *Inorganic Carbon Fluxes and Photosynthesis in Cyanobacteria – A Quantitative Model*. Netherlands: Progress in Photosynthesis Research, Springer, 289–293.
- Roussel, M. R., and Igamberdiev, A. U. (2011). Dynamics and mechanisms of oscillatory photosynthesis. *BioSystems* 103, 230–238. doi:10.1016/j.biosystems.2010.07.020
- Rovers, W., and Giersch, C. (1995). Photosynthetic oscillations and the interdependence of photophosphorylation and electron transport as studied by a mathematical model. *BioSystems* 35, 63–73. doi:10.1016/0303-2647(94)01481-L
- Rubio, F. C., Camacho, F. G., Sevilla, J. M. F., Chisti, Y., and Grima, E. M. (2003). A mechanistic model of photosynthesis in microalgae. *Biotechnol. Bioeng.* 81, 459–473. doi:10.1002/bit.10492
- Rügen, M., Bockmayr, A., and Steuer, R. (2015). Elucidating temporal resource allocation and diurnal dynamics in phototrophic metabolism using conditional FBA. *Sci. Rep.* 5, 15247. doi:10.1038/srep15247
- Rust, M. J., Golden, S. S., and O’Shea, E. K. (2011). Light-driven changes in energy metabolism directly entrain the cyanobacterial circadian oscillator. *Science* 331, 220–223. doi:10.1126/science.1197243
- Rust, M. J., Markson, J. S., Lane, W. S., Fisher, D. S., and O’Shea, E. K. (2007). Ordered phosphorylation governs oscillation of a three-protein circadian clock. *Science* 318, 809–812. doi:10.1126/science.1148596
- Saha, R., Versepunt, A. T., Berla, B. M., Mueller, T. J., Pakrasi, H. B., and Maranas, C. D. (2012). Reconstruction and comparison of the metabolic potential of cyanobacteria *Cyanothece* sp. ATCC 51142 and *Synechocystis* sp. PCC 6803. *PLoS ONE* 7:e48285. doi:10.1371/journal.pone.0048285
- Sakurai, I., Stazic, D., Eisenhut, M., Vuorio, E., Steglich, C., Hess, W. R., et al. (2012). Positive regulation of psbA gene expression by cis-encoded antisense RNAs in *Synechocystis* sp. PCC 6803. *Plant Physiol.* 160, 1000–1010. doi:10.1104/pp.112.202127
- Sauro, H. M. (2014). *Enzyme Kinetics for Systems Biology (Version 1.06)*, 2nd Edn. Ambrosius Publishing.
- Schaechter, M. (2015). A brief history of bacterial growth physiology. *Front. Microbiol.* 6:289. doi:10.3389/fmicb.2015.00289
- Schultze, M., Forberich, B., Rexroth, S., Dyczmons, N. G., Roegner, M., and Appel, J. (2009). Localization of cytochrome b6f complexes implies an incomplete respiratory chain in cytoplasmic membranes of the cyanobacterium *Synechocystis* sp. PCC 6803. *Biochim. Biophys. Acta* 1787, 1479–1485. doi:10.1016/j.bbabi.2009.06.010
- Schuermans, R. M., van Alphen, P., Schuermans, J. M., Matthijs, H. C. P., and Hellingwerf, K. J. (2015). Comparison of the photosynthetic yield of cyanobacteria and green algae: different methods give different answers. *PLoS ONE* 10:e0139061. doi:10.1371/journal.pone.0139061
- Shastri, A. A., and Morgan, J. A. (2005). Flux balance analysis of photoautotrophic metabolism. *Biotechnol. Prog.* 21, 1617–1626. doi:10.1021/bp050246d
- Shultzaberger, R. K., Boyd, J. S., Diamond, S., Greenspan, R. J., and Golden, S. S. (2015). Giving time purpose: the *Synechococcus elongatus* clock in a broader network context. *Annu. Rev. Genet.* 49, 485–505. doi:10.1146/annurev-genet-111212-133227
- Smith, R. M., and Williams, S. B. (2006). Circadian rhythms in gene transcription imparted by chromosome compaction in the cyanobacterium *Synechococcus elongatus*. *Proc. Natl. Acad. Sci. U.S.A.* 103, 8564–8569. doi:10.1073/pnas.0508696103
- Srinivasan, S., Cluett, W. R., and Mahadevan, R. (2015). Constructing kinetic models of metabolism at genome-scales: a review. *Biotechnol. J.* 10, 1345–1359. doi:10.1002/biot.201400522
- Steuer, R. (2007). Computational approaches to the topology, stability and dynamics of metabolic networks. *Phytochemistry* 68, 2139–2151. doi:10.1016/j.phytochem.2007.04.041
- Steuer, R., Gross, T., Selbig, J., and Blasius, B. (2006). Structural kinetic modeling of metabolic networks. *Proc. Natl. Acad. Sci. U.S.A.* 103, 11868–11873. doi:10.1073/pnas.0600013103
- Steuer, R., and Junker, B. H. (2009). “Computational models of metabolism: stability and regulation in metabolic networks,” in *Advances in Chemical Physics*, Vol. 142, ed. S. A. Rice (Hoboken, NJ: John Wiley & Sons, Inc).
- Steuer, R., Knoop, H., and Machné, R. (2012). Modelling cyanobacteria: from metabolism to integrative models of phototrophic growth. *J. Exp. Bot.* 63, 2259–2274. doi:10.1093/jxb/ers018
- Stöckel, J., Jacobs, J. M., Elvitigala, T. R., Liberton, M., Welsh, E. A., Polpitiya, A. D., et al. (2011). Diurnal rhythms result in significant changes in the cellular protein complement in the cyanobacterium *Cyanothece* 51142. *PLoS ONE* 6:e16680. doi:10.1371/journal.pone.0016680
- Stouthamer, A. H. (1973). A theoretical study on the amount of ATP required for synthesis of microbial cell material. *Antonie Van Leeuwenhoek* 39, 545–565. doi:10.1007/BF02578899
- Taheri-Araghi, S., Bradde, S., Sauls, J. T., Hill, N. S., Levin, P. A., Paulsson, J., et al. (2015). Cell-size control and homeostasis in bacteria. *Curr. Biol.* 25, 385–391. doi:10.1016/j.cub.2014.12.009
- Takai, N., Nakajima, M., Oyama, T., Kito, R., Sugita, C., Sugita, M., et al. (2006). A KaiC-associating SasA-RpaA two-component regulatory system as a major circadian timing mediator in cyanobacteria. *Proc. Natl. Acad. Sci. U.S.A.* 103, 12109–12114. doi:10.1073/pnas.0602955103
- Taniguchi, Y., Katayama, M., Ito, R., Takai, N., Kondo, T., and Oyama, T. (2007). labA: a novel gene required for negative feedback regulation of the cyanobacterial circadian clock protein KaiC. *Genes Dev.* 21, 60–70. doi:10.1101/gad.1488107
- Teng, S.-W., Mukherji, S., Moffitt, J. R., De Buyl, S., and O’Shea, E. K. (2013). Robust circadian oscillations in growing cyanobacteria require transcriptional feedback. *Science* 340, 737–740. doi:10.1126/science.1230996
- Terauchi, K., Kitayama, Y., Nishiwaki, T., Miwa, K., Murayama, Y., Oyama, T., et al. (2007). ATPase activity of KaiC determines the basic timing for circadian clock of cyanobacteria. *Proc. Natl. Acad. Sci. U.S.A.* 104, 16377–16381. doi:10.1073/pnas.0706292104
- Tomita, J., Nakajima, M., Kondo, T., and Iwasaki, H. (2005). No transcription-translation feedback in circadian rhythm of KaiC phosphorylation. *Science* 307, 251–254. doi:10.1126/science.1102540
- Tran, L. M., Rizk, M. L., and Liao, J. C. (2008). Ensemble modeling of metabolic networks. *Biophys. J.* 95, 5606–5617. doi:10.1529/biophysj.108.135442
- Tyystjärvi, E., Mäenpää, P., and Aro, E. M. (1994). Mathematical modelling of photoinhibition and photosystem II repair cycle. I. Photoinhibition and D1 protein degradation in vitro and in the absence of chloroplast protein synthesis in vivo. *Photosynth. Res.* 41, 439–449. doi:10.1007/BF02183046
- van Zon, J. S., Lubensky, D. K., Altena, P. R., and ten Wolde, P. R. (2007). An allosteric model of circadian KaiC phosphorylation. *Proc. Natl. Acad. Sci. U.S.A.* 104, 7420–7425. doi:10.1073/pnas.0608665104
- Vermaas, W. (2001). “Photosynthesis and respiration in cyanobacteria,” in *eLS* (John Wiley & Sons, Ltd).
- Vershubskii, A. V., Mishanin, V. I., and Tikhonov, A. N. (2014). Modeling of the photosynthetic electron transport regulation in cyanobacteria. *Biochem. Moscow Suppl. Ser. A* 8, 262. doi:10.1134/S199074781402007X
- Vijayan, V., Zuzov, R., and O’Shea, E. K. (2009). Oscillations in supercoiling drive circadian gene expression in cyanobacteria. *Proc. Natl. Acad. Sci. U.S.A.* 106, 22564–22568. doi:10.1073/pnas.0912673106
- Vu, T. T., Stolyar, S. M., Pinchuk, G. E., Hill, E. A., Kucek, L. A., Brown, R. N., et al. (2012). Genome-scale modeling of light-driven reductant partitioning and carbon fluxes in diazotrophic unicellular cyanobacterium *Cyanothece* sp. ATCC 51142. *PLoS Comput. Biol.* 8:e1002460. doi:10.1371/journal.pcbi.1002460
- Waldbauer, J. R., Rodrigue, S., Coleman, M. L., and Chisholm, S. W. (2012). Transcriptome and proteome dynamics of a light-dark synchronized bacterial cell cycle. *PLoS ONE* 7:e43432. doi:10.1371/journal.pone.0043432
- Wang, L., Birol, I., and Hatzimanikatis, V. (2004). Metabolic control analysis under uncertainty: framework development and case studies. *Biophys. J.* 87, 3750–3763. doi:10.1529/biophysj.104.048090

- Wijker, J. E., Jensen, P. R., Snoep, J. L., Vaz Gomes, A., Guiral, M., Jongsma, A. P., et al. (1995). Energy, control and DNA structure in the living cell. *Biophys. Chem.* 55, 153–165. doi:10.1016/0301-4622(94)00148-D
- Witzel, F., Götze, J., and Ebenhöf, O. (2010). Slow deactivation of ribulose 1,5-bisphosphate carboxylase/oxygenase elucidated by mathematical models. *FEBS J.* 277, 931–950. doi:10.1111/j.1742-4658.2009.07541.x
- Woelfle, M. A., Ouyang, Y., Phanvijhitsiri, K., and Johnson, C. H. (2004). The adaptive value of circadian clocks: an experimental assessment in cyanobacteria. *Curr. Biol.* 14, 1481–1486. doi:10.1016/j.cub.2004.08.023
- Woelfle, M. A., Xu, Y., Qin, X., and Johnson, C. H. (2007). Circadian rhythms of superhelical status of DNA in cyanobacteria. *Proc. Natl. Acad. Sci. U.S.A.* 104, 18819–18824. doi:10.1073/pnas.0706069104
- Wood, T. L., Bridwell-Rabb, J., Kim, Y.-I., Gao, T., Chang, Y.-G., LiWang, A., et al. (2010). The KaiA protein of the cyanobacterial circadian oscillator is modulated by a redox-active cofactor. *Proc. Natl. Acad. Sci. U.S.A.* 107, 5804–5809. doi:10.1073/pnas.0910141107
- Xu, M., Bernt, G., Singh, A., Mi, H., Rgner, M., Pakrasi, H. B., et al. (2008). Properties of mutants of *Synechocystis* sp. strain PCC 6803 lacking inorganic carbon sequestration systems. *Plant Cell Physiol.* 49, 1672–1677. doi:10.1093/pcp/pcn139
- Xu, Y., Mori, T., and Johnson, C. H. (2003). Cyanobacterial circadian clockwork: roles of KaiA, KaiB and the kaiBC promoter in regulating KaiC. *EMBO J.* 22, 2117–2126. doi:10.1093/emboj/cdg168
- Yang, Q., Pando, B. F., Dong, G., Golden, S. S., and van Oudenaarden, A. (2010). Circadian gating of the cell cycle revealed in single cyanobacterial cells. *Science* 327, 1522–1526. doi:10.1126/science.1181759
- Yao, D. C. I., Brune, D. C., and Vermaas, W. F. J. (2012). Lifetimes of photosystem I and II proteins in the cyanobacterium *Synechocystis* sp. PCC 6803. *FEBS Lett.* 586, 169–173. doi:10.1016/j.febslet.2011.12.010
- Yoda, M., Eguchi, K., Terada, T. P., and Sasai, M. (2007). Monomer-shuffling and allosteric transition in KaiC circadian oscillation. *PLoS ONE* 2:e408. doi:10.1371/journal.pone.0000408
- Yoshikawa, K., Aikawa, S., Kojima, Y., Toya, Y., Furusawa, C., Kondo, A., et al. (2015). Construction of a genome-scale metabolic model of *Arthrospira platensis* NIES-39 and metabolic design for cyanobacterial bioproduction. *PLoS ONE* 10:e0144430. doi:10.1371/journal.pone.0144430
- Zavřel, T., Červený, J., Knoop, H., and Steuer, R. (2016). Optimizing cyanobacterial product synthesis: meeting the challenges. *Bioengineered* 7, 490–496. doi:10.1080/21655979.2016.1207017
- Zhang, S., and Bryant, D. A. (2011). The tricarboxylic acid cycle in cyanobacteria. *Science* 334, 1551–1553. doi:10.1126/science.1210858
- Zhu, X.-G., de Sturler, E., and Long, S. P. (2007). Optimizing the distribution of resources between enzymes of carbon metabolism can dramatically increase photosynthetic rate: a numerical simulation using an evolutionary algorithm. *Plant Physiol.* 145, 513–526. doi:10.1104/pp.107.103713
- Zhu, X.-G., Govindjee, Baker, N. R., deSturler, E., Ort, D. O., and Long, S. P. (2005). Chlorophyll a fluorescence induction kinetics in leaves predicted from a model describing each discrete step of excitation energy and electron transfer associated with photosystem II. *Planta* 223, 114–133. doi:10.1007/s00425-005-0064-4
- Zhu, X.-G., Wang, Y., Ort, D. R., and Long, S. P. (2013). e-Photosynthesis: a comprehensive dynamic mechanistic model of C3 photosynthesis: from light capture to sucrose synthesis. *Plant Cell Environ.* 36, 1711–1727. doi:10.1111/pce.12025
- Zonneveld, C. (1998). Photoinhibition as affected by photoacclimation in phytoplakton: a model approach. *J. Theor. Biol.* 193, 115–123. doi:10.1006/jtbi.1998.0688
- Zwicker, D., Lubensky, D. K., and ten Wolde, P. R. (2010). Robust circadian clocks from coupled protein-modification and transcription–translation cycles. *Proc. Natl. Acad. Sci. U.S.A.* 107, 22540–22545. doi:10.1073/pnas.1007613107

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2016 Westermarck and Steuer. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

APPENDIX

The differential equations that define the model shown in **Figure 4** are $d[\text{PSU}^*]/dt = v_1 - v_2 - v_d$ and $d[\text{PSU}_d]/dt = v_d - v_r$, together with the conservation relationship $[\text{PSU}] + [\text{PSU}^*] + [\text{PSU}_d] = [\text{PSU}_{\text{total}}]$. The reactions rates describe the activation of the photosynthetic unit (PSU), $v_1 = k_1 \cdot I \cdot [\text{PSU}]$, the return to the open state, $v_2 = k_2 \cdot [\text{PSU}^*]$, the transition to a photodamaged

state, $v_d = k_d \cdot I \cdot [\text{PSU}^*]$, and the recovery, $v_r = k_r \cdot [\text{PSU}_d]$. The rate of activation, v_1 , and photodamage, v_d , both depend on the light intensity, I (with $k_d \ll k_1$). Parameters used are $k_1 = 10.0$, $k_2 = 10.0$, $k_r = 1.0$, $[\text{PSU}_{\text{total}}] = 1.0$, and k_d as indicated in **Figure 4**. The rate of oxygen evolution (v_2) obtained from an (analytical) steady-state solution of the ODEs is shown. Parameters are arbitrary and do not reflect actual values found in cyanobacteria.



Customized Steady-State Constraints for Parameter Estimation in Non-Linear Ordinary Differential Equation Models

Marcus Rosenblatt^{1*}, Jens Timmer^{1,2,3} and Daniel Kaschek¹

¹ Institute of Physics, Albert Ludwig University of Freiburg, Freiburg, Germany, ² Freiburg Centre for Systems Biology, Albert Ludwig University of Freiburg, Freiburg, Germany, ³ BIOS Centre for Biological Signaling Studies, Albert Ludwig University of Freiburg, Freiburg, Germany

OPEN ACCESS

Edited by:

Julio Vera González,
University Hospital Erlangen, Germany

Reviewed by:

Irene Otero Mulas,
Consejo Superior de Investigaciones
Científicas, Spain
Luis L. Fonseca,
Georgia Institute of Technology, USA

*Correspondence:

Marcus Rosenblatt
marcus.rosenblatt@fdm.uni-freiburg.de

Specialty section:

This article was submitted to
Systems Biology,
a section of the journal
Frontiers in Cell and Developmental
Biology

Received: 04 December 2015

Accepted: 21 April 2016

Published: 11 May 2016

Citation:

Rosenblatt M, Timmer J and
Kaschek D (2016) Customized
Steady-State Constraints for
Parameter Estimation in Non-Linear
Ordinary Differential Equation Models.
Front. Cell Dev. Biol. 4:41.
doi: 10.3389/fcell.2016.00041

Ordinary differential equation models have become a wide-spread approach to analyze dynamical systems and understand underlying mechanisms. Model parameters are often unknown and have to be estimated from experimental data, e.g., by maximum-likelihood estimation. In particular, models of biological systems contain a large number of parameters. To reduce the dimensionality of the parameter space, steady-state information is incorporated in the parameter estimation process. For non-linear models, analytical steady-state calculation typically leads to higher-order polynomial equations for which no closed-form solutions can be obtained. This can be circumvented by solving the steady-state equations for kinetic parameters, which results in a linear equation system with comparatively simple solutions. At the same time multiplicity of steady-state solutions is avoided, which otherwise is problematic for optimization. When solved for kinetic parameters, however, steady-state constraints tend to become negative for particular model specifications, thus, generating new types of optimization problems. Here, we present an algorithm based on graph theory that derives non-negative, analytical steady-state expressions by stepwise removal of cyclic dependencies between dynamical variables. The algorithm avoids multiple steady-state solutions by construction. We show that our method is applicable to most common classes of biochemical reaction networks containing inhibition terms, mass-action and Hill-type kinetic equations. Comparing the performance of parameter estimation for different analytical and numerical methods of incorporating steady-state information, we show that our approach is especially well-tailored to guarantee a high success rate of optimization.

Keywords: non-linear ODE models, parameter estimation, biochemical reaction networks, steady-state, positive solutions, multiplicity, multi-stability, success rate

1. INTRODUCTION

Dynamical systems are frequently modeled by systems of ordinary differential equations (ODEs). Homogeneously distributed molecules are treated as continuous quantities interacting with each other according to kinetic laws, e.g., mass-action or Michaelis-Menten kinetics.

A typical ODE system

$$\dot{x} = f(x, p, u(t)), \quad x(0) = x_0 \quad (1)$$

determines the time-evolution of an N -dimensional state vector $x(t)$. Here, $p \in \mathbb{R}_+^M$ denotes the M -dimensional vector of non-negative kinetic parameters. The vector $x_0 \in \mathbb{R}_{+,0}^N$, where $\mathbb{R}_{+,0} = \mathbb{R}_+ \cup \{0\}$, gives the set of initial conditions. The kinetic parameters and initial conditions together span the space of model parameters $\theta = (p, x_0)$. The explicit time-dependency via $u(t)$ corresponds to external driving forces, like drug stimuli in biological dynamic systems.

In many fields where ODE models are used, parameter values are not *a priori* known and have to be estimated from experimental data. Commonly, this is achieved by minimizing an objective function $g(\theta, D)$ that penalizes weighted differences between model prediction $x(t)$ and data D , e.g., by maximum-likelihood estimation. For the case of non-linear ODE systems, several local optima may exist. In order to find the global optimum, several optimization methods, e.g., particle swarm optimizers (Peng et al., 2010) or simulated annealing (Xiang and Gong, 2000), include stochasticity to escape local minima. Compared to that, deterministic algorithms may stick to local optima during optimization. On the other hand, gradient and Hessian information of the objective function can be incorporated, increasing the performance of optimization by a multiple. Combining the advantages of derivative-based optimization and random sampling, a multi-start deterministic optimization approach has proven to yield superior overall performance for our problem class (Raue et al., 2013). Throughout this work, we perform optimization by means of a trust-region optimizer from multiple starting positions.

Specially in models of biological systems, available data is sparse and parameters are often non-identifiable. Apart from that, the high-dimensional parameter space hampers parameter sampling. In order to reduce the number of parameters, the system is assumed to initially ($t = 0$) be in a steady-state which is determined by the constraint equation

$$f(x_0, p, 0) = 0. \quad (2)$$

As a standard approach, the steady-state constraint is solved for the initial values x_0 . Since Equation (2) is in general non-linear, this may lead to higher-order polynomial equations for which no general solution is available. Even for a rather simple case of quadratic or cubic equations, solutions are not unique and optimization would have to be performed for all possibilities. Another aspect of steady-state calculation are negative solutions for x_0 and p that appear for certain model specifications. Negative solutions are not only contradicting the biological setting with positively defined concentrations and kinetic parameters but also constitute a problem for optimization. Negative parameter values change the sign of damping terms of the ODE's right-hand side which might lead to rapidly growing solutions and an abort of the optimization before an optimum was reached.

In order to obtain a high convergence probability for the optimization of randomly chosen initial parameter samples, our

aim is to derive non-negative, analytical steady-state expressions, while multiple steady-state expressions are likewise circumvented by a proper choice of kinetic and initial value parameters for which Equation (2) is solved.

Over the last decades, steady-state analysis has been addressed by many algorithms and methods. In the following, we give an overview of existing approaches and summarize their applicability to different types of model equations with a special focus on parameter estimation in ODE models, see **Table 1**.

The earliest-proposed algorithm for deriving steady-states in enzyme-catalyzed systems being described by simple mass-action rules was developed by King and Altman (1956). In the original paper, however, interactions that do not involve the enzyme were not allowed which prohibits applicability to most of today's systems with proteins mediating the activation of other proteins without being part of the reaction. After Chemical Reaction Network Theory (CRNT) was formulated (Horn and Jackson, 1972; Feinberg, 1979), the method of King and Altman has been improved by graph theory (Chou, 1990) and extended to special subclasses of CRNs, e.g., layered signaling cascades (Feliu et al., 2012) and post-translational modification networks (Feliu and Wiuf, 2013). The same authors also published a more general approach for CRNs in Feliu and Wiuf (2012). Here, a set of core variables is introduced serving for a parametrization of the steady-states whereby non-negative solutions are guaranteed due to graph-theoretical arguments.

Another approach developed by Halasz et al. (2013), introduces bilinearities of the system as new variables leading to a linearized system solvable by application of Cramer's rule. The number of bilinearities, however, is restricted and negative steady-state solutions are not prevented.

All mentioned approaches deal with steady-state analysis for CRNs based on mass-action rules. However, modern modeling approaches often make use of special reaction types such as inhibition, Michaelis-Menten or Hill kinetics that cannot be included into standard CRNT without changing the model structure and introducing new dynamical variables. In the approach of Halasz et al. (2013), inhibition and Michaelis-Menten terms can easily be integrated by multiplying the corresponding steady-state equation by the denominator of the rate expression. However, since a state variable is contained in the denominator, this can increase the number of bilinearities significantly.

TABLE 1 | Covered reaction types and positivity for different methods of steady-state determination.

| Method | CRN, Mass action, Production, Degradation | Inhibition, Michaelis-Menten | Hill Kinetic | Non-negative Solutions |
|----------------------------------|---|------------------------------|--------------|------------------------|
| King-Altman | Yes | No | No | Yes |
| Feliu and Wiuf | Yes | No | No | Yes |
| Halasz et al. | Yes | Yes | No | No |
| Loriaux et al. (py-substitution) | Yes | Yes | Yes | No |
| Proposed | Yes | Yes | Yes | Yes |

In order to avoid problems of higher-order polynomial equations, steady-state equations can be solved not only for initial value- but also for kinetic parameters, which is done in the steady-state solver *py-substitution* developed by Loriaux et al. (2013). From N initial values and M kinetic parameters, a set of N variables is chosen that have to be fixed in Equation (2). In doing so, a lot of freedom is incorporated into the solution. In fact, *py-substitution* is able to solve the very most steady-state equation systems, since in principle N kinetic parameters could be chosen as fixed variables directly leading to a simple linear equation system.

Complementary to analytical approaches, steady-state information can be incorporated into the system by numerically computing the initial conditions during each optimization step. Even gradient information that is necessary for efficient optimization is available by means of the implicit function theorem. A numerical incorporation of steady-state information has the advantage that the complexity of the underlying equation system is in principal not restricted. Furthermore, the implementation remains untouched when model equations are changed. However, convergence of the numerical steady-state calculation is not guaranteed and issues of multiple steady-states cannot be controlled.

In the following section, we present a method to derive non-negative steady-state expressions for a large class of nonlinear ODE models that are based on biochemical reactions. Our approach picks up the idea of solving for kinetic parameters in order to derive unique and simple steady-state expressions. Due to the structure of the ODE system, solving for kinetic parameters often leads to potentially negative steady-state solutions, depending on the point of evaluation in parameter space. By introducing appropriate parameter transformations and exploiting the given model structure, our approach guarantees a non-negative solution space. In the Results section, we show how different steady-state parameterizations influence the optimization procedure and compare our approach to the

standard approach of solving for initial value parameters as well as to a numerical steady-state approach.

2. METHODS

2.1. Theoretical Background

Let us consider a *model* f as an N -dimensional ODE system $\dot{x} = f(x, p)$ with states x , parameters p and no external driving forces, i.e., the ODE is autonomous. We write f as a matrix product

$$f(x, p) = S \cdot F(x, p), \quad (3)$$

of the $N \times M$ -dimensional *stoichiometry matrix* S and the M -dimensional *flux vector* F which depends on states and parameters. For the entries of the flux vector, we allow rational functions of x and p including e.g., mass-action, inhibition, Michaelis-Menten and Hill-Type kinetics. **Table 2** gives an overview of the main reaction types covered by the presented steady-state approach.

We assume that each single flux F_l is proportional to some *flux parameter* k_l and can be written as

$$F_l = k_l \cdot G_l(x, q), \quad (4)$$

where the function G_l only depends on the states and a set of *additional parameters* q taken from the set of all model parameters p . The union of flux parameters k and additional model parameters q coincides with the parameter set p . Typically all reaction types described by CRNT only need one flux parameter and do not contribute to q , however, inhibition terms and Michaelis-Menten kinetics contain at least one additional parameter and Hill kinetics even two.

The signs of the entries of the stoichiometry matrix S determine whether a flux contributes as an *in*- or an *outflux* to the time evolution of the corresponding state. We assume that each outflux is at least linearly dependent on the corresponding

TABLE 2 | Examples of typical reaction types of ODE models.

| # | Stoichiometry | Reaction type | Flux | Contribution to $\frac{dA}{dt}$ |
|----|------------------------------|-----------------------------|---|---------------------------------|
| 1 | $\emptyset \rightarrow A$ | Production | k | Positive/influx |
| 2 | $A \rightarrow \emptyset, B$ | Degradation, transformation | $k \cdot A$ | Negative/outflux |
| 3 | $A + A \rightarrow AA$ | Dimerization | $k \cdot A^2$ | Negative/outflux |
| 4 | $B \rightarrow A$ | Transformation | $k \cdot B$ | Positive/influx |
| 5 | $B + C \rightarrow A$ | Binding | $k \cdot B \cdot C$ | Positive/influx |
| 6 | $B \rightarrow A$ | Inhibition by C | $k \cdot B \cdot \frac{1}{q+C}$ | Positive/influx |
| 7 | $B \rightarrow A$ | Michaelis-Menten | $k \cdot B \cdot \frac{1}{q+B}$ | Positive/influx |
| 8 | $A + B \rightarrow C$ | Binding | $k \cdot A \cdot B$ | Negative/outflux |
| 9 | $A \rightarrow B$ | Inhibition by C | $k \cdot A \cdot \frac{1}{q+C}$ | Negative/outflux |
| 10 | $A \rightarrow B$ | Michaelis-Menten | $k \cdot A \cdot \frac{1}{q+A}$ | Negative/outflux |
| 11 | $B \rightarrow A$ | Hill | $k \cdot \frac{B^{q_2}}{q_1^{q_2} + B^{q_2}}$ | Positive/influx |
| 12 | $\emptyset \rightarrow A$ | Self-activation | $k \cdot \frac{A^{q_2}}{q_1^{q_2} + A^{q_2}}$ | Positive/influx |
| 13 | $B + C \rightarrow A$ | Power-law | $k \cdot B^{q_1} \cdot C^{q_2}$ | Positive/influx |

All types are covered by our steady-state approach.

state, as being always the case for mass-action systems. By means of Equation (2), each initial value $x_{0,i} \equiv x_i$ is directly related with a steady-state equation of the form

$$0 = \sum \text{in}_i - x_i \cdot \sum \text{out}_i, \quad (5)$$

where in_i and out_i constitute functions of states and parameters. For a majority of reaction types used in ODE models, the fluxes in_i and out_i are independent of x_i , compare **Table 2**. In these cases, Equation (5) is linear in x_i and has the solution $x_i = \frac{\sum \text{in}_i}{\sum \text{out}_i}$. However, if the fluxes in_i or out_i still depend on x_i , e.g., reaction 12 in **Table 2** for the case of self-activation or reaction 3 with an outflux being quadratic in x_i , Equation (5) might be non-linear in x_i .

In order to solve the complete set of steady-state equations, we analyze their specific structure by means of graph theory. We therefore rewrite Equation (5) as

$$x_i = \frac{\sum \text{in}_i}{\sum \text{out}_i} \quad (6)$$

and summarize appearances of states on the right-hand side of Equation (6). Here, the set of states is defined by the set of dynamic variables x that we want to fix by the steady-state determination. Once a dynamic variable is fixed by a non-negative expression or treated as a free parameter, it is removed from the set of states.

Definition 1: A *head* of state x_i is a state x_j that appears on the right-hand side of Equation (6). By $h(x_i)$, we refer to the set of heads for a specific state x_i . In particular, x_i can itself be a part of $h(x_i)$.

Proposition 1: If non-negative steady-state solutions for all heads of x_i are known, a non-negative steady-state solution for x_i can directly be obtained by Equation (6). This holds especially, if the set $h(x_i)$ is empty.

Definition 2: The *adjacency matrix* $M(f)$ of an ODE model $f(x, p)$ with states x and parameters p is an $N \times N$ matrix with entries

$$M_{ji} = \begin{cases} 1, & \text{if } x_j \in h(x_i) \\ 0, & \text{else.} \end{cases}$$

Each d_M -dimensional adjacency matrix M defines a *directed graph* G_M with nodes x_1 to x_{d_M} which we call *steady-state graph*. Each non-zero entry of M corresponds to a directed edge (x_j, x_i) implying that x_j occurs in the steady-state expression of x_i , i.e., Equation (6). A non-zero diagonal entry M_{ii} reflects that the corresponding steady-state equation is non-linear in x_i .

2.2. Splitting Cycles

The specific structure of the steady-state graph enables to solve the steady-state equations step-by-step as is shown in the following.

Definition 3: A *cycle* of a steady-state graph is a path through the graph along its edges with equal starting and end point. Here, we allow cycles of length one arising from non-zero diagonal entries in the adjacency matrix M .

Definition 4: Graphs that do not contain cycles are called *tree-like*.

Proposition 2: If a steady-state graph of an N -dimensional model f is tree-like, non-negative steady-state solutions can be obtained for all x_i inside the graph.

Proof: For any tree-like steady-state graph, there exists at least one *root*, i.e., state without head, called x_r . Since $h(x_r) = \emptyset$, the corresponding steady-state expression can be obtained by *Proposition 1*. In doing so, x_r is removed from the steady-state graph and a new state serves as root for which *Proposition 1* again gives the corresponding steady-state expression. By iteratively applying *Proposition 1* for each of the N nodes, the complete steady-state solution is obtained.

Considering *Proposition 2*, it is clear that solving the steady-state constraint Equation (2) for the set of initial values only becomes intricate, if there are cycles inside the steady-state graph such that higher-order polynomial equations arise. The idea of our steady-state approach is to split all these cycles step-by-step such that *Proposition 2* can ultimately be applied to the remaining graph.

The simplest way of splitting a cycle is by means of a conserved quantity (CQ) of the system arising from the stoichiometry. A general introduction can be found in Loriaux et al. (2013) or Halasz et al. (2013). The following definition restricts to the properties being relevant for the presented approach.

Definition 5: A *conserved quantity* (CQ) of the model f is an expression of states and parameters which remains constant during the time-evolution of f . For each CQ, the number of independent steady-state equations is reduced by one implying that one state or flux parameter that appears in the CQ can be chosen freely. If all CQs can be derived from the stoichiometry matrix, the number of CQs is given by $n_{cq} = N - R_S$, with the model size N and the rank of the stoichiometry matrix R_S . The cases for which $n_{cq} > N - R_S$ are discussed in Section 2.5.

In order to split a cycle by a CQ, one of the states, x_c , that appears both inside the cycle and in the CQ is chosen freely. The corresponding steady-state equation is removed from Equation (2), whereby the number of independent steady-state equations remains constant. Since the state x_c is treated as a free variable, all edges originating from and leading to x_c can be removed from the steady-state graph and the considered cycle is split. Note, that each CQ can only be used once.

If no states inside the cycle appear in CQs, the cycle can be split by solving the steady-state equation of a specific cycle state x_i for a flux parameter k_i . By means of Equations (3) and (4), the steady-state expression of k_i holds

$$k_i = \frac{-1}{S_{ii}G_i(x, q)} \sum_{j \neq i} S_{ij}k_j G_j(x, q). \quad (7)$$

Proposition 3: Let n_{k_i} be the number of appearances of the flux parameter k_i , see Equation (4), inside the steady-state constraint Equation (2). Then n_{k_i} coincides with the number of non-zero entries in the i -th column of the stoichiometry matrix S .

Unless k_i does not appear in other steady-state equations, i.e., $n_{k_i} = 1$, the considered cycle is removed from the steady-state graph without affecting other parts of the graph. However, if

$n_{k_l} > 1$, all further appearances have to be substituted by Equation (7) which creates new edges inside the steady-state graph and possibly even new cycles. In order to keep the structure as simple as possible, flux parameters with $n_{k_l} = 1$ play a special role.

2.3. Enforcing Positivity

Although solving for flux parameters implies linear equations and therefore structurally simple steady-state expressions, the solutions are often negative for certain model specifications. Here, we show how positivity of the expressions can be guaranteed by appropriate transformations.

The steady-state expression, Equation (7), of the flux parameter k_l was derived by means of the steady-state equation of x_i . The expression contains minus signs if and only if at least one of the stoichiometry entries $S_{i,j \neq l}$ has the same sign as S_{il} . In this case, at least one further flux contributes to x_i with the same sign as $F_l = k_l G_l$, namely as an in- or outflux.

Definition 6: For the steady-state equation of x_i , Equation (5), we define μ_i and ν_i as the number of in- and outfluxes, respectively. Furthermore, we define the *dimension* of the state x_i as the minimum $\dim(x_i) = \min(\mu_i, \nu_i)$.

If $\dim(x_i) = 1$, a non-negative steady-state expression is obtained by solving for the particular flux parameter being the only in- or outflux, compare first examples in **Table 3**.

If $\dim(x_i) > 1$, positivity can be enforced by performing an appropriate parameter transformation. In order to do so, we divide the fluxes contributing to x_i into influxes $F_{in,1} \dots F_{in,\mu_i}$ and outfluxes $F_{out,1} \dots F_{out,\nu_i}$. Then, Equation (5) reads

$$0 = \sum_{j=1}^{\mu_i} F_{in,j} - \sum_{l=1}^{\nu_i} F_{out,l}. \quad (8)$$

Let us assume that we want to solve Equation (8) for the influx parameter $k_{in,1} = \frac{F_{in,1}}{G_{in,1}}$. We perform a variable transformation by defining the ratio between the remaining influxes and $F_{in,1}$ as

$$r_z = \frac{F_{in,z}}{F_{in,1}} = k_{in,z} \cdot \frac{G_{in,z}}{k_{in,1} G_{in,1}} \quad \text{for } z = 2, \dots, \mu_i, \quad (9)$$

TABLE 3 | Examples of solving steady-state equations for flux parameters.

| Steady-state equation | Solve for | Type | Solution |
|---|------------|------|--|
| $0 = k_1 A^2 - k_2 X$ | k_1 | 1 | $k_1 = \frac{k_2 X}{A^2}$ |
| $0 = k_1 A^2 - k_2 X$ | k_2 | 1 | $k_2 = \frac{k_1 A^2}{X}$ |
| $0 = k_1 A^2 - k_2 X - k_3 X B - \frac{k_4 X}{q_1 + C}$ | k_1 | 1 | $k_1 = \frac{X}{A^2} \left(k_2 + k_3 B + \frac{k_4}{q_1 + C} \right)$ |
| $0 = k_1 A^2 + k_5 B - k_2 X - k_3 X B - \frac{k_4 X}{q_1 + C}$ | k_1, k_5 | 2 | $k_1 = \frac{1}{A^2} X \left(k_2 + k_3 B + \frac{k_4}{q_1 + C} \right) \cdot \frac{1}{1+r}$ $k_5 = \frac{1}{B} X \left(k_2 + k_3 B + \frac{k_4}{q_1 + C} \right) \cdot \frac{r}{1+r}$ |

where the r_z replace the kinetic parameters k_2 to k_{μ_i} . By means of Equations (8) and (9), we obtain

$$\sum_l F_{out,l} = F_{in,1} \cdot \left(1 + \sum_{j=2}^{\mu_i} r_{in,j} \right)$$

and therefore

$$k_{in,1} = \frac{1}{G_{in,1}} \sum_l F_{out,l} \cdot \frac{1}{1 + \sum_{j=2}^{\mu_i} r_j}. \quad (10)$$

Since $G_{in,1}$ and $F_{out,l}$ are positive and Equation (10) is a sum of positive contributions, a non-negative steady-state expression for $k_{in,1}$ is guaranteed. By means of Equation (9), the remaining flux parameters have to be substituted by the non-negative expressions

$$k_{in,z} = \frac{1}{G_{in,z}} \sum_l F_{out,l} \cdot \frac{r_z}{1 + \sum_{j=2}^{\mu_i} r_j} \quad \text{for } z = 2, \dots, \mu_i. \quad (11)$$

For an outflux parameter, we analogously obtain

$$k_{out,1} = \frac{1}{G_{out,1}} \sum_j F_{in,j} \cdot \frac{1}{1 + \sum_{l=1}^{\nu_i} r_l} \quad \text{and} \quad (12)$$

$$k_{out,z} = \frac{1}{G_{out,z}} \sum_j F_{in,j} \cdot \frac{r_z}{1 + \sum_{l=1}^{\nu_i} r_l} \quad \text{for } z = 2, \dots, \nu_i. \quad (13)$$

2.4. Algorithm for Steady-State Determination

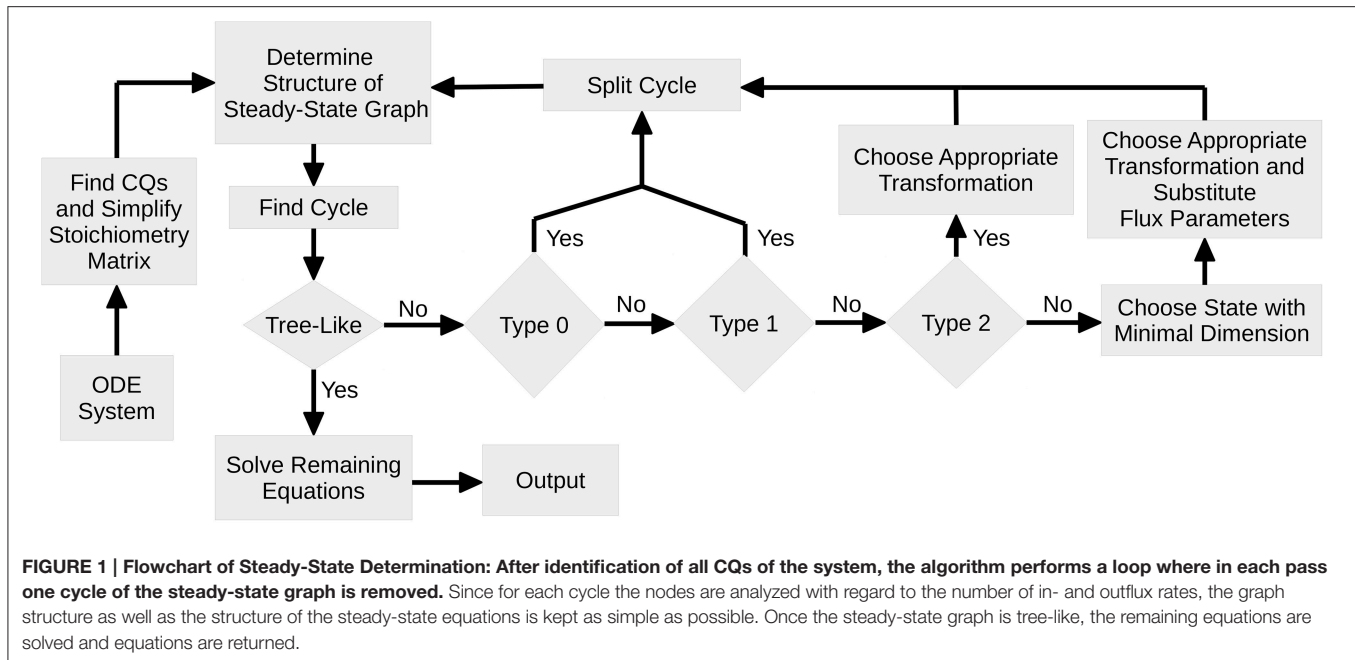
In the previous sections, we showed how simple steady-state expressions can be obtained (Section 2.2), while positivity is likewise guaranteed (Section 2.3). In order to split one cycle of the steady-state graph and solve for a flux parameter, a pair (x_i, k_j) of state and flux parameter has to be chosen, which is not unique. In the following, we suggest an algorithm based on a classification of such pairs.

According to Definitions 5, 6 and Proposition 3, we associate each pair with one of four different types:

$$(x_i, k_j) \equiv \begin{cases} \text{Type 0, if } x_i \text{ appears in a CQ} \\ \text{Type 1, } n_{k_j} = 1 \text{ and } \dim(x_i) = 1 \\ \text{Type 2, } n_{k_j} = 1 \text{ and } \dim(x_i) > 1 \\ \text{Type 3, else.} \end{cases}$$

Figure 1 shows a flowchart of the algorithm. At first, the set of CQs is computed for the ODE system serving as an input for the algorithm. If the graph is tree-like, the remaining equations are obtained according to *Proposition 2* and the complete set of steady-state equations is returned.

In the case of a pair of Type 0, the cycle can simply be removed by interpreting the corresponding state as a free variable. The CQ that is thereby used is removed from the set of CQs and cannot



further contribute to the steady-state determination. Here, the flux parameters remain unaffected.

Unless the cycle cannot be directly split by means of a CQ, the corresponding steady-state equation, Equation (5), is solved for one of the $\mu_i + v_i$ flux parameters by use of a pair of Type 1, 2 or 3. In order to keep the steady-state solution as simple as possible, pairs of Type 1 are preferred, since this enables to split the cycle both without substituting the flux parameter by its steady-state expression, Equation (7), and without introducing flux ratios as new parameters, Equation (9).

If no pairs of Type 1 are available, the algorithm scans the steady-state graph for pairs of Type 2. In this case a parameter transformation is necessary in order to guarantee positivity of the solution. However, the flux parameter does not appear any more in the system and therefore has not to be substituted. In all three cases, Type 0, 1, or 2, the number of cycles of the steady-state graph is reduced.

If pairs of Type 2 are also not available, all pairs are of Type 3. In this case, it is not *a priori* clear which pair is the best choice. As a simply revisable choice, the algorithm then solves the steady-state equation of the state with minimal dimension. Subsequently, all further appearances of flux parameters have to be replaced by their particular transformation, Equations (11) or (13).

2.5. Calculating the Conserved Quantities and Simplifying the Stoichiometry Matrix

In order to find CQs of the ODE system, linear combinations of rows of the stoichiometry matrix S can be analyzed. According to Equation (3), the N -dimensional ODE system can be written as

$$\dot{x} = S \cdot F(x, p). \quad (14)$$

Multiplication of Equation (14) by an $N \times N$ -matrix M yields

$$M \cdot \dot{x} = \tilde{S} \cdot F(x, p), \quad (15)$$

where the matrix $\tilde{S} = M \cdot S$ defines linear combinations of rows of S . For each row \tilde{S}_i that is equal to zero, the quantity $M_i \cdot x = \sum_j M_{ij} x_j$ is conserved.

In fact, each set of linearly dependent rows of S implies a CQ. For some ODE systems, however, not all CQs can be derived from S without accounting for the flux vector F . Equation (14) can be written

$$\dot{x} = C(p, x) \cdot x. \quad (16)$$

where $C(p, x)$ is an $N \times N$ -matrix dependent on the parameters and states. Analyzing linear dependencies of C , all CQs of the form

$$\sum_j a_j(x, p) \cdot x_j = \text{const.} \quad (17)$$

can be found, where the coefficients a_j might depend on states and parameters.

In order to determine symbolic expressions for the a_j , we transpose the matrix C and numerically search for linearly dependent columns. All parameters and states appearing in C^T are replaced by random values to obtain a numeric matrix C_{ran}^T for which a QR -decomposition is performed. The matrix R constitutes an upper triangular matrix, where the number of non-empty rows corresponds to the rank of C_{ran} . The first column R_ℓ with $R_{\ell\ell} = 0$ is a linear combination of the columns $R_{j<\ell}$. Therefore, also the column C_ℓ is a linear combination of the columns $C_{j<\ell}$ implying that the equation system $\sum_j a_j C_j = 0$

has a solution for the a_j with $a_{j>\ell} = 0$ which can be calculated symbolically. Thus, the quantity $a^T \cdot x$ is conserved. Once a CQ has been found, one of the corresponding linearly dependent rows of the stoichiometry matrix is removed and the procedure is repeated. In most ODE systems, all CQs of the system can be obtained in that way. For all other cases, our `Python` code provides a possibility to manually specify CQs.

The idea of taking linear combinations of the stoichiometry matrix S , see Equations (14) and (15), can be augmented to simplify S for the calculation of steady-state expressions. For each matrix M , the original steady-state constraint, $S \cdot F = 0$, is replaced by a new set of steady-state equations, $\tilde{S} \cdot F = 0$. With a clever choice of M , these new steady-state equations might be structurally simpler than the original ones. With respect to our proposed algorithm, the matrix M should (1) minimize the overall number of entries in the new stoichiometry matrix \tilde{S} and (2) prevent the creation of new cycles. In practice, the idea of linearly combining rows of the stoichiometry matrix can lead to structurally simpler steady-state expressions as we show by means of a small example in the Supplementary Material.

2.6. Numerically Computed Steady-States

Besides calculating steady-states analytically, roots of the steady-state constraint, Equation (2), can be computed numerically during each step of the optimization. Here, we perform Newton's method which is fast compared to the time of the ODE integration. The gradient information that is necessary within our deterministic optimization scheme is determined by the implicit function theorem, i.e., given the steady-state constraint

$$f(x_0(p), p) = 0,$$

we derive the equation with respect to p and obtain

$$0 = \frac{\partial f}{\partial x_0} \cdot \frac{\partial x_0}{\partial p} + \frac{\partial f}{\partial p} \implies \frac{\partial x_0}{\partial p} = - \left(\frac{\partial f}{\partial x_0} \right)^{-1} \frac{\partial f}{\partial p}.$$

2.7. Technical Remarks

The steady-state algorithm was implemented in `Python` by use of the libraries `numpy` and `sympy`. It can either be downloaded from the author's homepage as a `Python` code or can be used from within the R-packages `dMod/cOde` available from <https://github.com/dkaschek/>. Simulation of data and parameter estimation with analytical and numerical steady-states were performed in `dmod`.

3. RESULTS

When calculating steady-state expressions for parameter estimation of ODE systems, several aspects have to be considered simultaneously. Most importantly, the parameter space is to be reduced as far as possible. Therefore, all available steady-state constraints should be taken into account. Since solving for state variables often leads to higher-order equations for which solutions are difficult to obtain, one has at least partially to solve for kinetic parameters. In doing so, the steady-state expressions

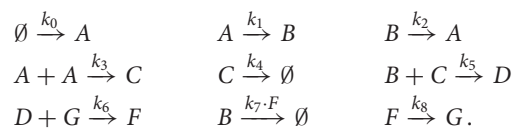
often lead to negative parameter values for certain model specifications.

Due to mass balance, outfluxes contribute with a minus sign to the time derivative of the corresponding state. Provided that outflux rates are proportional to positive powers of their states, they contribute damping terms to the time-evolution of the state. However, if for a certain model specification the corresponding flux parameter is negative, the sign of the outflux term becomes positive which leads to an exploding model trajectory for the state.

In Section 3.1, we show how our steady-state approach determines simple steady-state equations for systems that lead to higher-order equations when solved for the state variables. In Section 3.2, we show how steady-state expressions with negative realizations lead to optimization problems and a significantly lower success rate, i.e., the probability to converge to a local or the global optimum. Non-linear ODE systems often have several steady-state solutions, when the steady-state equations are solved for the state variables. For parameter estimation, multiple steady-states constitute a problem, since all possible realizations have in principle to be followed up. By solving for kinetic parameters, our steady-state approach likewise avoids multiple solutions and improves the optimization as we show in Section 3.3.

3.1. Determination of Non-Negative Steady-State Expressions

To show the applicability of the presented steady-state approach, we investigate a toy model with six state variables and nine reactions of the form



All reactions satisfy the law of mass action, the degradation of B is mediated by F . With these assumptions, one obtains the following ODE system

$$\begin{aligned} \dot{A} &= k_0 + k_2B - k_1A - k_3A^2 \\ \dot{B} &= k_1A - k_2B - k_5BC - k_7BF \\ \dot{C} &= k_3A^2 - k_5BC - k_4C \\ \dot{D} &= k_5BC - k_6DG \\ \dot{F} &= k_6DG - k_8F \\ \dot{G} &= k_8F - k_6DG. \end{aligned}$$

The system contains one conserved quantity $F + G = \text{const}$, reflecting that the steady-state equations of F and G are not independent from each other. Therefore, the number of variables that have to be fixed by the steady-state is five. In order to obtain the corresponding steady-state equations, all time-derivatives of the states are set to zero. Although the single equations of this system are of degree two or lower, solving for the states leads to a sixth order polynomial equation, see Supplementary Material, for which no closed-form solution is available.

TABLE 4 | Steps of steady-state determination for a model with six states and eight reactions.

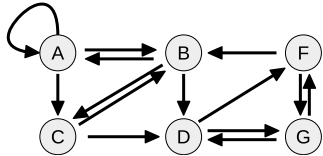
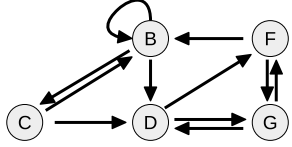
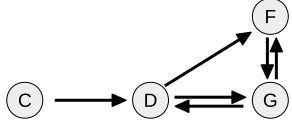
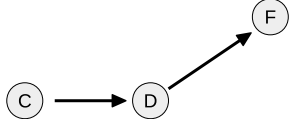
| Loop | Steady-state graph | Steady-state equations | Simplification |
|------|---|--|--|
| 1 |  | $A: 0 = k_0 + k_2B - k_1A - k_3A^2$ $B: 0 = k_1A - k_2B - k_5BC - k_7BF$ $C: 0 = k_3A^2 - k_5BC - k_4C$ $D: 0 = k_5BC - k_6DG$ $F: 0 = k_6DG - k_8F$ $G: 0 = k_8F - k_6DG$ | Cycle: [A, A] Type 3 $k_0 = \frac{A^2k_3 + Ak_1}{r_1 + 1}$ $k_2 = r_1 \frac{k_0}{B} = r_1 \frac{A^2k_3 + Ak_1}{B(r_1 + 1)}$ |
| 2 |  | $B: 0 = k_1A \frac{1}{r_1 + 1} - k_3A^2 \frac{r_1}{r_1 + 1} - k_5BC - k_7BF$ $C: 0 = k_3A^2 - k_5BC - k_4C$ $D: 0 = k_5BC - k_6DG$ $F: 0 = k_6DG - k_8F$ $G: 0 = k_8F - k_6DG$ | Cycle: [B, B] Type 1 $k_1 = \frac{r_1 + 1}{A} [k_3A^2 \frac{r_1}{r_1 + 1} + k_5BC + k_7BF]$ |
| 3 |  | $C: 0 = k_3A^2 - k_5BC - k_4C$ $D: 0 = k_5BC - k_6DG$ $F: 0 = k_6DG - k_8F$ $G: 0 = k_8F - k_6DG$ | Cycle: [D, G, D] Type 0 G is part of CQ |
| 4 |  | $C: 0 = k_3A^2 - k_5BC - k_4C$ $D: 0 = k_5BC - k_6DG$ $F: 0 = k_6DG - k_8F$ | Tree-like Solve for C, D and F |

Table 4 summarizes how our steady-state solver determines a non-negative steady-state solution by partially solving for flux parameters. During the first loop, the cycle [A, A] of state A to itself is split. The pairs of A and its contributing flux parameters are all of Type 3, since there are two influx- and two outflux parameters of which at least one is appearing in the other steady-state equations, e.g., in the equation of B. The equation of A is solved for the influx parameter k_0 , whereby k_2 is transformed and replaced by the new free parameter $r_1 = k_2B/k_0$, see first loop in **Table 4**. The appearance of k_2 in the equation of B is substituted, whereas k_0 has no further appearances.

Whereas the state A is removed from the steady-state graph, the state B has become a new head of itself in consequence of the substitution. In the second loop, this new cycle [B, B] is split by solving the equation of B for the flux parameter k_1 . Here, the pair (B, k_1) is of Type 1, since k_1 is the only influx parameter and not appearing in the remaining equations.

In the next loop, the algorithm splits the cycle [D, G, D] by taking G as a free parameter, since it is part of the conserved quantity $F + G$. The remaining steady-state graph in the last loop is tree-like and therefore the steady-state equations can be derived according to *Proposition 2* starting with C which in this case serves as the root of the graph.

For simplification of writing, our steady-state solver outputs the equations in a specific order where fixed states or parameters may still appear in the equations below. In order to obtain a complete independent set of equations, one has to replace step by step. For the presented example, the ultimately obtained expressions are

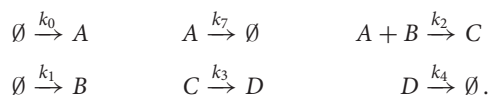
$$\begin{aligned}
 F &= A^2 \frac{k_3}{k_8} \frac{Bk_5}{Bk_5 + k_4} \\
 D &= A^2 \frac{k_3}{Gk_6} \frac{Bk_5}{Bk_5 + k_4} \\
 C &= A^2 \frac{k_3}{Bk_5 + k_4} \\
 k_1 &= Ak_3r_1 + \frac{r_1 + 1}{A} (BCK_5 + BFk_7) \\
 k_2 &= r_1 \frac{k_3A^2 + k_1A}{B(r_1 + 1)} \\
 k_0 &= \frac{k_3A^2 + k_1A}{r_1 + 1},
 \end{aligned}$$

where six parameters are fixed, while one additional parameter r_1 can be chosen freely.

3.2. Minus Signs Imply a Low Convergence Rate

For a given data set and a given ODE model, each parameter set determines the time-evolution of the states and its likelihood L can be computed based on the data. Here, parameter values are estimated by minimizing the negative log-likelihood function $-\log L$. For the case of non-linear ODE models, several local optima may exist. In order to find the global optimum, we perform multi-start optimization in combination with a trust-region optimizer. A powerful optimization approach should have a high probability to find a local or the global optimum.

Let us consider an ODE system with four state variables and six reactions of the form



The corresponding ODE system is given by

$$\begin{aligned} \dot{A} &= k_0 - k_7 A - k_2 AB \\ \dot{B} &= k_1 - k_2 AB \\ \dot{C} &= k_2 AB - k_3 C \\ \dot{D} &= k_3 C - k_4 D. \end{aligned}$$

In order to test if negative steady-state expressions lead to optimization problems, we implemented four different steady-state parameterizations, see **Table 5**, and compared the success rate of parameter optimization. For each approach, six parameters are optimized as shown in **Table 5**. Besides the standard approach, i.e., exclusively solving for initial values, two other parameterizations were derived by solving the equation of state *B* for two different kinetic parameters, namely k_7 and k_0 . The latter guarantees a non-negative steady-state solution. Apart from that, a fourth parameterization was constructed by adding the equation $k_0 = k_1 + \Delta_{k_0}$ to the standard steady-state formulation. In doing so, k_0 is transformed such that $k_0 > k_1$, with the new free parameter Δ_{k_0} describing the difference between k_0 and k_1 . This approach likewise implies positivity.

TABLE 5 | Several steady-state representations for a model with four states and six flux parameters.

| Steady-state method | Steady-state equations | Parameters to be estimated |
|---------------------------------------|--|---|
| Solved for initial values | $A = k_0 / (k_2 B + k_7)$ $B = k_1 k_7 / (k_2 (k_0 - k_1))$ $C = k_1 / k_3$ $D = k_1 / k_4$ | k_1, k_2, k_3, k_4 |
| (Standard) | | k_0, k_7 |
| Equation of <i>B</i> solved for k_7 | $A = k_0 / (k_2 B + k_7)$ $k_7 = B k_2 (k_0 - k_1) / k_1$ $C = k_1 / k_3$ $D = k_1 / k_4$ | k_1, k_2, k_3, k_4 k_0, B |
| Equation of <i>B</i> solved for k_0 | $A = k_0 / (k_2 B + k_7)$ $k_0 = k_1 k_7 / (B k_2) + k_1$ $C = k_1 / k_3$ $D = k_1 / k_4$ | k_1, k_2, k_3, k_4 B, k_7 |
| Standard | $k_0 = k_1 + \Delta_{k_0}$ $A = k_0 / (k_2 B + k_7)$ $B = k_1 k_7 / (k_2 (k_0 - k_1))$ $C = k_1 / k_3$ $D = k_1 / k_4$ | k_1, k_2, k_3, k_4 Δ_{k_0}, k_7 |

For simulation of data, we chose a set of kinetic parameters for ODE integration, initialized the system with its steady-state and excited it by displacement of *A* at time point $t = 30$. Data points were generated for 16 different time points by adding normally distributed noise to the model trajectories. In order to study a scenario with different experimental conditions, i.e., different stimulations, simulation was done for three different displacement values, compare *cond1*, *cond2*, and *cond3* in **Figure 2A**.

Figure 2A shows data points and trajectories of a model fit that reached the global optimum. For each steady-state parameterization, we optimized 200 different parameter samples and counted how often several optima were reached. In **Figure 2B**, all converged fits are shown in order of the objective value, in our case the negative log-likelihood value. Several steps corresponding to local optima appear for all steady-state parameterizations, the deepest step corresponds to the global optimum. It can be concluded that the two parameterizations without minus signs, i.e., *Solving for k_0* and *Standard with Trafo*, show a significantly better convergence than the other two. For example, in the parameterization with transformation, the global optimum was twice as often reached than in the standard approach.

In order to explain the convergence behavior of the different steady-state implementations, we analyzed the correlation between initial parameter guess and the success of optimization. The steady-state of the presented model is negative, if and only if $k_0 < k_1$. **Figure 2C** shows the starting samples along the parameter axes of k_0 and k_1 for all four steady-state parameterizations, colors indicate whether a sample did not converge (black) or did converge to a local (blue) or the global optimum (yellow). For comparison, **Figure 2D** shows starting samples along axes of parameters that do not affect the sign of the steady-state, namely k_2 and k_4 . The sample distribution shows that samples with $k_0 > k_1$ have a high probability to converge, while samples with $k_0 < k_1$ tend to abort. On the other hand, the relation of k_2 and k_4 does not have a significant impact on the convergence probability. Furthermore, **Figure 2C** shows that the reparameterized steady-states prohibit sampling in the region with $k_0 < k_1$.

In addition to the starting samples, we analyzed the parameter paths during the optimization. **Figures 2E,F** show the paths for the first 50 starting samples with respect to the above used parameter axes. Parameter samples with $k_0 < k_1$ usually abort without any considerable steps in parameter space even though several samples cross the border $k_0 = k_1$ and proceed. In the opposite direction, some samples reach the border when started in the area with $k_0 > k_1$ and abort exactly at the border. The very most samples drawn with $k_0 > k_1$ converged to a local or the global optimum *G*. Again, **Figure 2F** underlines that the convergence behavior is unaffected by the relation of k_2 and k_4 .

We conclude that steady-state parameterizations that lead to negative parameter values for certain model specifications constitute a severe issue for optimization. Due to the formulation of our steady-state algorithm, negative solutions are automatically avoided in the obtained steady-state expressions.

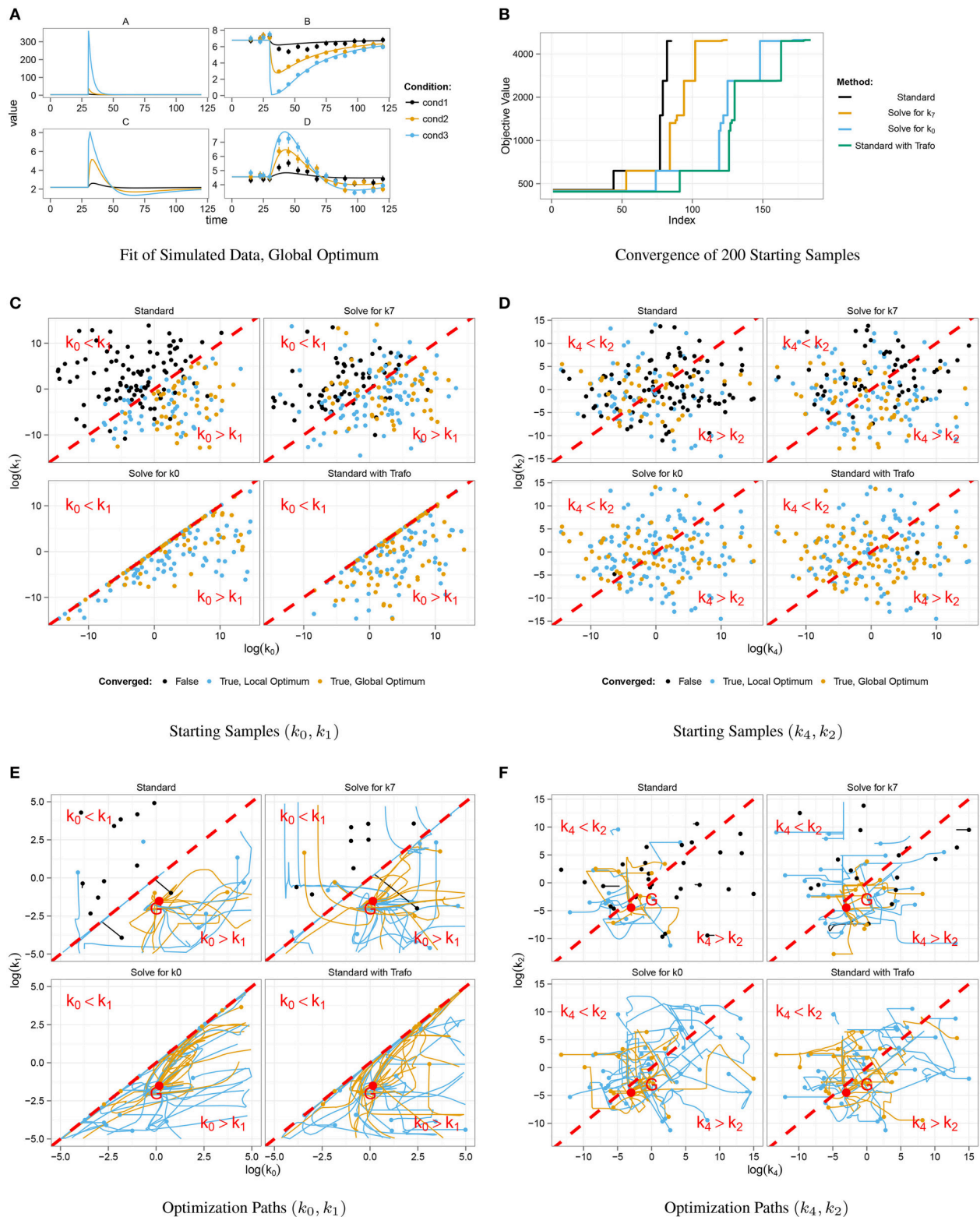
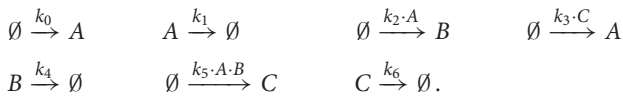


FIGURE 2 | Optimization results for different steady-state parameterizations. Data was simulated for three different displacements of A at $t = 30$ (A).

Convergent fits for all four steady-state implementations were sorted by increasing objective value (B). Steps correspond to local minima. Positive steady-state parameterizations show a considerably better convergence behavior. Starting samples are shown in different colors in (C,D), indicating whether the corresponding optimization converged. Parameter paths starting with $k_0 < k_1$ did mostly not converge as opposed to samples with $k_0 > k_1$ which mostly converged to a local or the global optimum (E,F).

3.3. Dealing with Multiplicity of Steady-States

Let us consider a system with three state variables and seven reactions of the form



The production of B is mediated by A , production of A is mediated by C and production of C is mediated by both A and B . The corresponding ODE system is given by

$$\begin{aligned} \dot{A} &= k_0 + k_3 C - k_1 A \\ \dot{B} &= k_2 A - k_4 B \\ \dot{C} &= k_5 AB - k_6 C. \end{aligned} \quad (18)$$

For this system, the steady-state can still be analytically solved for the states A , B and C . The solution reads

$$\begin{aligned} C_{1/2} &= \frac{1}{2k_2 k_3 k_5} \left(\frac{1}{k_1^2 k_4 k_6} \Delta + 2k_0 k_2 k_3 k_5 \pm \sqrt{\Delta} \right) \\ B_{1/2} &= \frac{1}{2k_1 k_3 k_4 k_5} \left(k_1^2 k_4 k_6 \pm \sqrt{\Delta} \right) \\ A_{1/2} &= \frac{1}{2k_1 k_2 k_3 k_5} \left(k_1^2 k_4 k_6 \pm \sqrt{\Delta} \right), \end{aligned} \quad (19)$$

with the discriminant $\Delta = k_1^2 k_4 k_6 \cdot (k_1^2 k_4 k_6 - 4k_0 k_2 k_3 k_5)$. For $\Delta > 0$, two positive steady-state solutions $S_1 = (A_1, B_1, C_1)$ and $S_2 = (A_2, B_2, C_2)$ are obtained, while the system has no real steady-state for $\Delta < 0$.

Linear stability analysis reveals that solution S_1 is unstable and solution S_2 is stable, see Supplementary Material. If several steady-state solutions exist, only one of them can be chosen for an optimization run at a time. Here, the stable solution was chosen.

For this system, the issue of multiplicity can easily be solved, however, for more complicated systems several stable solutions might exist. Stable solutions might even switch to unstable solutions along the optimization path, e.g., in case of a Hopf bifurcation. For higher-order equations, analytical solutions become unfeasible and numerical steady-state computation comes into play. However, since the numerical root finding is performed by means of Newton's method, the result depends on initial guesses for A , B and C . Consequently, it is not clear which of the solutions is obtained and stability of the retrieved steady-state is not guaranteed. As we will show, coexistence of stable and unstable steady-state solutions leads to a reduced convergence probability in the numerical approach.

Unlike solving the steady-state equations for A , B and C , the steady-state expressions obtained by our proposed approach are

$$C = A^2 \frac{k_2 k_5}{k_4 k_6} \quad B = A \frac{k_2}{k_4} \quad k_1 = A \frac{k_2 k_3 k_5}{k_4 k_6} + \frac{k_0}{A}, \quad (20)$$

where the kinetic parameter k_1 is fixed, while the initial value of A is taken as a free parameter. The obtained solution is unique, since the steady-state equations are linear in the parameters B , C and k_1 .

In general, our approach avoids multiple-steady-states by choosing a combination of parameters for which the steady-state equations are linear. In doing so, no solution is neglected as long as all steady-state equations are fulfilled. As an analogon, let us consider a single algebraic equation of the form $ab^3 + cb^2 + db + f = 0$, with the five parameters a , b , c , d and f . On the one hand this equation can be solved for b whereby multiple solutions are obtained. On the other hand it can be solved for one of the parameters a , c , d or f for which the equation is linear leading to a unique solution.

In the following, we compare the convergence behavior of three different steady-state implementations, namely *Standard*, i.e., analytically solved for the states, *Numeric*, i.e., numerically solved for the states, and *Proposed*, i.e., our steady-state approach with positive solutions. For the former two implementations, the seven kinetic parameters k_0 to k_6 are estimated, whereas for the proposed approach, the initial value of A is estimated instead of k_1 , compare Equation (20).

Since natural systems are always subject to external noise, unstable steady-states are never realized by the system. Therefore, data was simulated by means of the stable steady-state solution. Analogously to Section 3.2, three different displacements of the state A were triggered at time point $t = 30$ to excite the system. Here, data points were generated for eight different time points.

In order to test the convergence behavior, we started 200 fits from randomly chosen parameter samples. **Figure 3A** shows an example fit that converged to the global optimum. The optimization result of the three steady-state approaches is compared in **Figure 3B**. Steps correspond to local optima. In our approach, nearly half of the samples converged to the global optimum, whereas only about 10% of the fits converged in the standard approach and even less in the numeric approach.

Similar to Section 3.2, we analyzed the correlation between initial parameter guess and success in optimization. **Figure 3D** shows the distribution of starting samples with respect to the sign of the discriminant Δ . For $\Delta < 0$ the discriminant of the standard steady-state expression, Equation (19), becomes negative, and all starting samples drawn from this region did not converge. In addition, **Figure 3E** shows that the optimization of these samples directly aborted, since the path did not take any or at most a very small step in parameter space.

Furthermore, we analyzed the correlation between the coexistence of stable and unstable steady-state solutions and the success of the numerical approach. During each optimization step, the root of the ODE's right-hand side is computed for the current parameter values. Depending on the initial guess, either the stable or the unstable solution is obtained. In order to see, if the unstable solution causes optimization aborts, we chose state A as a representative and compared numerically and analytically calculated values at the end of the optimization

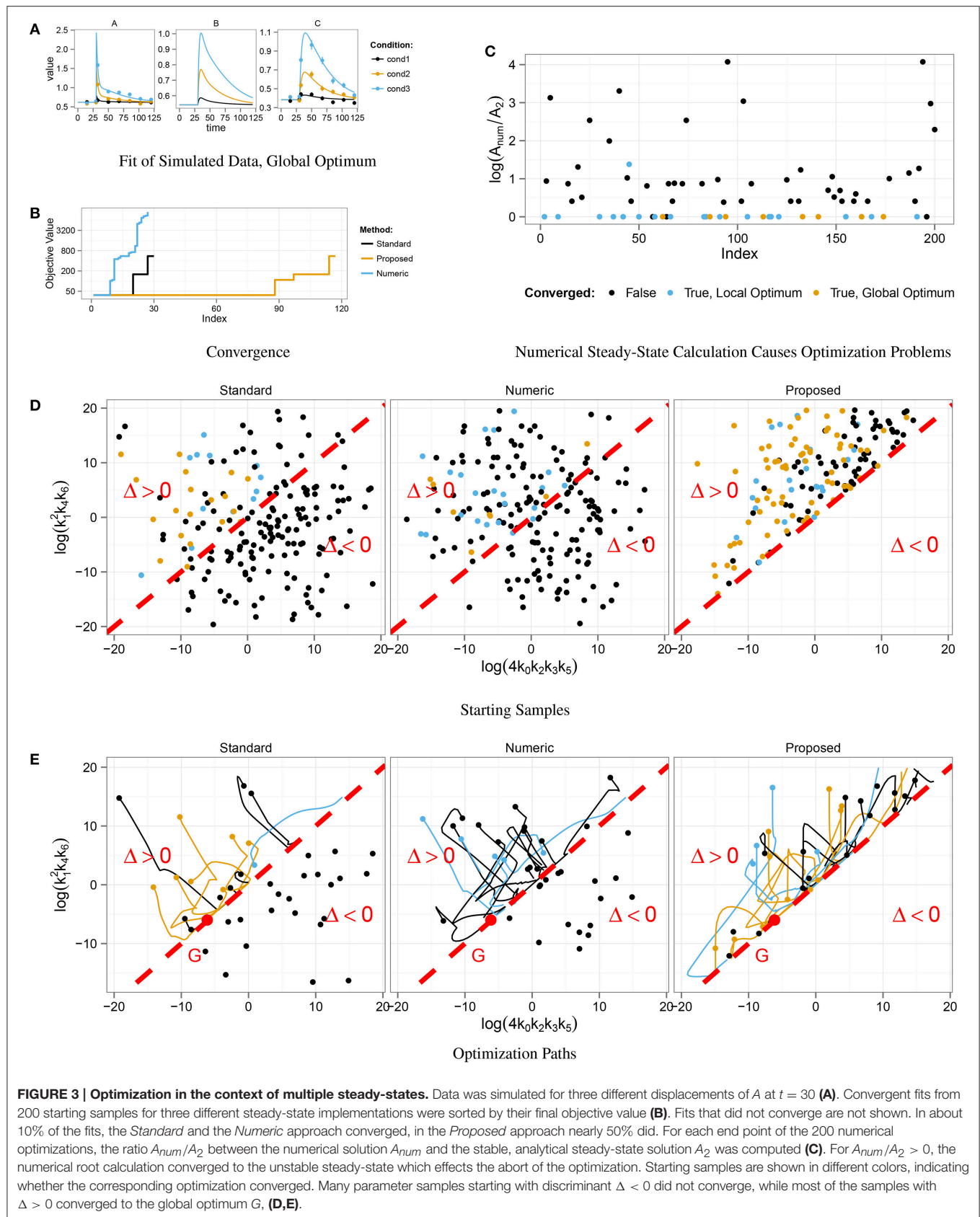


FIGURE 3 | Optimization in the context of multiple steady-states. Data was simulated for three different displacements of A at $t = 30$ (A). Convergent fits from 200 starting samples for three different steady-state implementations were sorted by their final objective value (B). Fits that did not converge are not shown. In about 10% of the fits, the *Standard* and the *Numeric* approach converged, in the *Proposed* approach nearly 50% did. For each end point of the 200 numerical optimizations, the ratio A_{num}/A_2 between the numerical solution A_{num} and the stable, analytical steady-state solution A_2 was computed (C). For $A_{num}/A_2 > 0$, the numerical root calculation converged to the unstable steady-state which effects the abort of the optimization. Starting samples are shown in different colors, indicating whether the corresponding optimization converged. Many parameter samples starting with discriminant $\Delta < 0$ did not converge, while most of the samples with $\Delta > 0$ converged to the global optimum G, (D,E).

path. The numerically calculated value A_{num} was taken from the root calculation by Newton's method and the value A_2 of the stable steady-state was calculated by means of Equation (19) with the corresponding parameter values. **Figure 3C** shows ratios A_{num}/A_2 for all fits. If $A_{num}/A_2 = 0$, the stable steady-state was obtained, while $A_{num}/A_2 > 0$ implies that the unstable solution was obtained. Since nearly all fits that reached the unstable solution did not converge, we conclude that the coexistence of a second unstable steady-state causes optimization aborts in the numerical approach of steady-state determination.

Both problems arising from the existence of multiple steady-states, i.e., negative discriminants and stable vs. unstable steady-states, are automatically circumvented by our steady-state algorithm resulting in a superior convergence rate during parameter estimation.

4. DISCUSSION AND CONCLUSION

Parameter estimation in non-linear ODE models of biological systems has to deal with several local optima and a high-dimensional parameter space. In order to reduce the number of parameters, steady-state constraints are taken into account. Deterministic algorithms search for the global optimum by performing the optimization with multiple starting samples. The way of implementing steady-states, i.e., the exact parameterization, has an impact on the convergence probability of a randomly chosen starting sample. If optimizations tend to abort before reaching an optimum, many starting samples are necessary to find the best possible fit. Since incorporation of steady-state information shifts parameter distributions and contributes to gradient information, the exact steady-state parametrization plays a crucial role in optimization.

For many systems, steady-state equations lead to higher-order polynomial equations when being solved for the state variables. To exploit the full steady-state information, equations can be partially solved for kinetic parameters. If the obtained steady-state expressions yield negative values for certain parameter specifications, those might lead to rapidly growing solutions for the ODE system. We showed that negative parameter values have a considerable, negative impact on the success of the optimization.

In many applications, multiplicity and multi-stability of the steady-state constitutes the relevant question. In the case of parameter estimation, however, multiple steady-states complicate the estimation process. For the standard approach of solving steady-state equations for the state variables, all solutions principally have to be considered and optimization has to be performed for all possibilities. For the numerical implementation of steady-states also unstable steady-state solutions constitute a problem, since the numerical root finding method might converge to the unstable solution. In our case, the convergence probability dropped by 80%.

In this work, we presented an algorithm that derives steady-state expressions and circumvents negative and multiple solutions by construction. The approach covers the most common classes of ODE models consisting of e.g., mass-action kinetics, inhibition terms, Michaelis-Menten or Hill-type equations. By means of graph theory, cyclic dependencies between dynamical variables, e.g., positive or negative feedbacks inside a signaling cascade that lead to polynomial equations of order two or higher are removed by solving for kinetic parameters for which the equations are linear. In order to guarantee positivity of all solutions, the algorithm performs appropriate parameter transformations replacing kinetic parameters by ratios of participating fluxes. Our approach experiences a major limitation if simultaneously, the size of the ODE model becomes large and combinations of several in- and outflux parameters contribute to multiple states. Then, the algorithm is not able to find a strictly positive solution for the system. Furthermore, since the algorithm solves for rate parameters, it might not be applicable if solving for parameters is not allowed due to other reasons, e.g., if rate parameters must take certain fixed values.

In summary, our approach enables steady-state calculation for models with many cyclic dependencies that lead to higher-order polynomial equations when solved for state variables. Multiplicity and multi-stability are avoided and positivity of the solution is guaranteed. The parameter space is reduced by the number of independent steady-state equations while the nice convergence behavior is preserved.

AUTHOR CONTRIBUTIONS

MR developed, implemented and tested the method. MR, JT, and DK together wrote the paper.

FUNDING

This work was supported by the Federal Ministry of Education and Research (BMBF), by the MIP-DILI project, Innovative Medicines Initiative Joint Undertaking under grant agreement No. 115336-2 and by the ReelinSys project, Systems biology of Reelin-associated neuropsychiatric disorders, under No. 0316174D.

ACKNOWLEDGMENTS

We thank all our colleagues who provided us with many test models that contributed to the development of the algorithm.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: <http://journal.frontiersin.org/article/10.3389/fcell.2016.00041>

REFERENCES

- Chou, K.-C. (1990). Applications of graph theory to enzyme kinetics and protein folding kinetics. *Biophys. Chem.* 35, 1–24. doi: 10.1016/0301-4622(90)80056-D
- Feinberg, M. (1979). Complex balancing in general kinetic systems. *Arch. Rat. Mech. Anal.* 49, 187–194. doi: 10.1007/BF00255665
- Feliu, E., Knudsen, M., Andersen, L. N., and Wiuf, C. (2012). An algebraic approach to signaling cascades with n layers. *Bull. Math. Biol.* 74, 45–72. doi: 10.1007/s11538-011-9658-0
- Feliu, E., and Wiuf, C. (2012). Variable elimination in chemical reaction networks with mass-action kinetics. *SIAM J. Appl. Math.* 72, 959–981. doi: 10.1137/110847305
- Feliu, E., and Wiuf, C. (2013). Variable elimination in post-translational modification reaction networks with mass-action kinetics. *J. Math. Biol.* 66, 281–310. doi: 10.1007/s00285-012-0510-4
- Halász, A., Lai, H.-J., McCabe Pryor, M., Radhakrishnan, K., and Edwards, J. (2013). Analytical solution of steady-state equations for chemical reaction networks with bilinear rate laws. *IEEE/ACM Trans. Comput. Biol. Bioinform.* 10, 957–969. doi: 10.1109/TCBB.2013.41
- Horn, F., and Jackson, R. (1972). General mass action kinetics. *Arch. Rat. Mech. Anal.* 47, 81–116. doi: 10.1007/BF00251225
- King, E. L., and Altman, C. (1956). A schematic method of deriving the rate laws for enzyme-catalyzed reactions. *J. Phys. Chem.* 60, 1375–1378. doi: 10.1021/j150544a010
- Loriaux, P. M., Tesler, G., and Hoffmann, A. (2013). Characterizing the relationship between steady state and response using analytical expressions for the steady states of mass action models. *PLoS Comput. Biol.* 9:e1002901. doi: 10.1371/journal.pcbi.1002901
- Peng, H., Li, L., Yang, Y., and Liu, F. (2010). Parameter estimation of dynamical systems via a chaotic ant swarm. *Phys. Rev. E* 81:016207. doi: 10.1103/PhysRevE.81.016207
- Raue, A., Schilling, M., Bachmann, J., Matteson, A., Schelker, M., Kaschek, D., et al. (2013). Lessons learned from quantitative dynamical modeling in systems biology. *PLoS ONE* 8:e74335. doi: 10.1371/journal.pone.0074335
- Xiang, Y., and Gong, X. G. (2000). Efficiency of generalized simulated annealing. *Phys. Rev. E* 62, 4473–4476. doi: 10.1103/PhysRevE.62.4473

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2016 Rosenblatt, Timmer and Kaschek. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



Time Hierarchies and Model Reduction in Canonical Non-linear Models

Hannes Löwe, Andreas Kremling and Alberto Marin-Sanguino *

Specialty Division for Systems Biotechnology, Technische Universität München, Garching, Germany

The time-scale hierarchies of a very general class of models in differential equations is analyzed. Classical methods for model reduction and time-scale analysis have been adapted to this formalism and a complementary method is proposed. A unified theoretical treatment shows how the structure of the system can be much better understood by inspection of two sets of singular values: one related to the stoichiometric structure of the system and another to its kinetics. The methods are exemplified first through a toy model, then a large synthetic network and finally with numeric simulations of three classical benchmark models of real biological systems.

OPEN ACCESS

Edited by:

Ioannis P. Androulakis,
Rutgers University, USA

Reviewed by:

Preetam Ghosh,
Virginia Commonwealth University,
USA
Spyros K. Stamatelos,
Icahn School of Medicine at Mount
Sinai, USA

*Correspondence:

Alberto Marin-Sanguino
a.Marin@lrz.tu-muenchen.de

Specialty section:

This article was submitted to
Systems Biology,
a section of the journal
Frontiers in Genetics

Received: 27 February 2016

Accepted: 05 September 2016

Published: 21 September 2016

Citation:

Löwe H, Kremling A and
Marin-Sanguino A (2016) Time
Hierarchies and Model Reduction in
Canonical Non-linear Models.
Front. Genet. 7:166.
doi: 10.3389/fgene.2016.00166

Keywords: mathematical modeling, biochemical systems theory, quasi-polynomial systems, time-scales, model reduction, systems biology

1. INTRODUCTION

Biochemical systems are amenable to be modeled using differential equations but, due to the great diversity of mechanisms involved, the resulting models lack a defined structure. There is seldom a common set of properties that might simplify their analysis or enable the development of general tools. Models with a well defined structure enable a great level of abstraction and generality. Control engineering using linear systems is a case in point, where chemical plants, steam engines, and electric systems can all be treated within the same framework. In opposition to that, the analysis of *ad-hoc* biological models is often restricted to the numerical integration of a few scenarios.

The difficulties to analyze biological systems start with identification of which components to include in—or exclude from—the model, since the cellular milieu contains many, highly interconnected components. In addition to that, the intervening processes often progress at different time-scales, the resulting models tend to be stiff and difficult to analyze. But multiple time-scales also offers an opportunity for a deeper analysis (Hek, 2010). Many properties of biochemical systems are tied to the time hierarchy of the system. For instance, a regulatory mechanism must be as fast or faster than the process it is supposed to stabilize while some types of oscillations come from the interaction between a fast and a slow subsystem. An analysis of the first case need only take into account the subsystem that corresponds to the right time-scale, while the second case would better be analyzed by focusing on interactions between a fast and a slow subsystem. Furthermore, separating time scales reduces the stiffness of the system, and the computing power needed for numerical integration of the models.

A wide variety of time scale separation methods is available (Gerdtzen et al., 2004; Jamshidi and Palsson, 2008) but no single all-round solution has been found due to the difficulties associated to the diversity of biological models and their non-linearity.

Many model reduction methods are based in rewriting the model—e.g., through non-dimensionalization—in a form in which the different time-scales are shown explicitly

$$\begin{aligned}\dot{x} &= \varepsilon f(x, y, t), & x(0) &= x_0 \\ \dot{y} &= g(x, y, t), & y(0) &= y_0\end{aligned}\quad (1)$$

since the derivative of x is multiplied by the small perturbation parameter ε , it will have slower dynamics than y . This is a regular perturbation problem for which an approximate solution can be obtained by writing the equations when $\varepsilon \rightarrow 0$. The solution of:

$$\begin{aligned}\dot{x} &= 0, & x(0) &= x_0 \\ \dot{y} &= g(x, y, t), & y(0) &= y_0\end{aligned}\quad (2)$$

describes the fast dynamics of the system for intervals of time small enough that the change in x is negligible. The solution to this simplified problem is called inner solution and it is generally valid for the thin time slice $t = \mathcal{O}(\varepsilon)$. In order to obtain the slow dynamics, a reparametrization $\tau = \varepsilon t$ can be performed to obtain:

$$\begin{aligned}\frac{dx}{d\tau} &= f(x, y, \tau), & x(0) &= x_0 \\ \varepsilon \frac{dy}{d\tau} &= g(x, y, \tau), & y(0) &= y_0\end{aligned}\quad (3)$$

Which is a singular perturbation problem, since the order of the equations changes when $\varepsilon \rightarrow 0$, yielding.

$$\begin{aligned}\frac{dx}{d\tau} &= f(x, y, \tau), & x(0) &= x_0 \\ 0 &= g(x, y, \tau)\end{aligned}\quad (4)$$

Since the original differential equation for y becomes an algebraic equation, its initial condition $y(0) = y_0$ can no longer be satisfied. However, provided that the eliminated equation for y had a hyperbolic solution (one lacking a central manifold), this approximation will be valid for $t \gg \varepsilon$ and it is also known as the outer solution.

These kinds of problem fall within the category known as boundary layer problems, alluding to the transition between the inner and the outer solution. Obtaining a uniformly valid solution for all times, requires the matching of the inner and outer solution, however, when one is interested in the behavior of the system well within the area of validity of each solution, as is the case in biology, the inner and outer solutions are informative enough.

The appearance of algebraic equations introduces difficulties of its own. When the solution of Equation (2) can be written explicitly, $y = \phi(x, t)$, then algebraic constants can be eliminated by back-substitution in the o.d.e.s:

$$\dot{x} = f(x, \phi(x, t), t), \quad x(0) = x_0 \quad (5)$$

Finding this solution and substituting it as well as the non-dimensionalization step itself are no easily accomplished for big non-linear systems. The wide variety of possible structures for the equations is a challenge for any attempt to do this systematically.

2. MATERIALS AND METHODS

2.1. Modal Analysis

The advantage of dealing with a system that has a regular, convenient structure is made evident by analyzing time scales in the linear case. It has been shown (Palsson et al., 1987; Jamshidi and Palsson, 2008) that linearizing around a certain steady state and decomposing the Jacobian matrix of the system, allows to define aggregate variables or modes:

$$\mathbf{J} = \mathbf{M}^{-1} \mathbf{\Lambda} \mathbf{M} \quad (6)$$

where $\mathbf{\Lambda}$ is a diagonal matrix with the eigenvalues of \mathbf{J} . Some eigenvalues/eigenvectors may be imaginary conjugates, in that case, a similar decomposition may be used where $\mathbf{\Lambda}$ will be a Jordan canonical form. In any case, new variables can be defined:

$$\mathbf{m} = \mathbf{M} \mathbf{x} \quad (7)$$

and the linearized differential equation would be:

$$\dot{\mathbf{m}} = \mathbf{\Lambda} \mathbf{m} \quad (8)$$

Since $\mathbf{\Lambda}$ is diagonal, each mode m_i will vary independently of the rest in the linearized system, except for modes corresponding to conjugate pairs of eigenvalues, which will remain bound together. Modes enable to find combinations of variables with different timescales even for cases when the time scales of all the variables are similar. Modes work ideally with linear systems since the modes themselves are linear combinations of the variables. Back-substituting linear expressions in a linear system does not alter its structure, because of the telescopic property. For a non-linear system, however, the Jacobian matrix changes in every point so the modes will only be uncoupled at the point where the system is linearized. furthermore, non-linear systems do not normally comply with the above mentioned telescopic property, which results in differential-algebraic systems.

In subsequent sections we will apply these and similar concepts to a very general class systems that, in spite of being non-linear, have a regular structure and some extremely useful properties.

2.2. Canonical Non-linear Forms

The theoretical results of this work arise from the properties of the power-law and quasi-polynomial formalisms. These two formalisms have been shown to be mathematically equivalent. Whether a model is simpler in one mechanism or another depends on the particular processes involved. In general, power-law models are preferred to describe processes that depend on absolute fluxes (e.g., chemical networks) and quasi-polynomial models are used for modeling processes based on *per capita* rates, like logistic equations or classical predator-prey models.

In any case, any system of differential equations that fulfills one of the formalisms can be rewritten to the other without any loss of information. Furthermore, virtually any non-linear system of differential equations can be rewritten as one of the above mentioned systems through approximation (Savageau, 1969a), detailed mechanistic representation (Savageau, 1998), exact recasting (Savageau and Voit, 1987), or partitioning of its parameter space (Savageau et al., 2009; Lomnitz and Savageau, 2015).

2.2.1. Power-Law

The most general expression for these models is as linear combinations of rates or fluxes:

$$\dot{\mathbf{x}} = \mathbf{N} \mathbf{v}(\mathbf{x}) \quad (9)$$

where each term $v_i = \gamma_i \prod_j x_j^{f_{i,j}}$ are the power-laws the formalism takes the name from. The rate constants, γ_i are positive real numbers and the kinetic orders, $f_{i,j}$, are real numbers normally between -2 and 2 . It is also common to include “inputs” to the system as the so called independent variables, that reflect the environment in which the system operates and remain constant during a simulation or experiment. These variables can be included as part of γ without loss of generality. This kind of model is called Generalized Mass Action (GMA). All the formalisms discussed here can be expressed in a very convenient form adopting a direct notation proposed by Lewis Voit (1991) and that is slowly being adopted for theoretical analyses involving power-laws (Marin-Sanguino et al., 2010; Müller et al., 2016). So GMA equations becomes:

$$\dot{\mathbf{x}} = \mathbf{N} \text{diag}(\gamma) \mathbf{x}^{\mathbf{F}} \quad (10)$$

where the notation $\text{diag}(\cdot)$ will be used to represent a diagonal matrix containing vector (\cdot) as its main diagonal. All the information on the system is summarized in two matrices and a vector: \mathbf{N} of size $n \times m$ reflects the stoichiometry of the system—mass conversion/conservation— \mathbf{F} has size $m \times n$ and contains the kinetic information. The $m \times 1$ sized vector, γ , serves as a reference connecting rates and metabolites—e.g., when the system variables are normalized by their value at a certain equilibrium point, $z_i = x_i/|x|_0$, then the vector of rate constants becomes the vector of steady state fluxes of the system. Under such conditions, the partition of information becomes clear between a stoichiometric/static-flux information $\mathbf{N} \text{diag}(\gamma)$ and kinetics \mathbf{F} a particular type of gma models, the s-systems, have received an exceptional deal of attention due to their remarkable properties. An s-system has a single positive and a single negative term:

$$\dot{\mathbf{x}} = \text{diag}(\alpha) \mathbf{x}^{\mathbf{G}} - \text{diag}(\beta) \mathbf{x}^{\mathbf{H}} \quad (11)$$

where α and β are rate constants and \mathbf{G} and \mathbf{H} are kinetic order matrices. these systems have analytic solutions for their steady states (Savageau, 1969b).

The variables in a s-system can be normalized using their steady state values (Savageau, 1974). Defining new variables $z_i = \frac{x_i}{|x_i|_0}$ where the zero subindex indicates the numerical value of the variable in the steady state, and rearranging terms, results in the system:

$$\dot{\mathbf{z}} = \text{diag}(\mathbf{f}) (\mathbf{z}^{\mathbf{G}} - \mathbf{z}^{\mathbf{H}}) \quad (12)$$

Due to this normalization, the new variables will reach the steady state at $z_i = 1 \forall i$. The factors f_i , are the turnovers of their respective variables at the steady-state (Savageau, 1974),

$$f_i = \left| \frac{V_i^+}{x_i} \right|_0 = \left| \frac{V_i^-}{x_i} \right|_0 \quad (13)$$

and considered to contain information relative to the time scale of the corresponding variable. Actually F -values are the reciprocals of transition times as defined by Easterby (1981).

2.2.2. Quasi-Polynomial

In their more general form, quasi-polynomial systems can be written as Generalized Lotka–Volterra (GLV)

$$\dot{\mathbf{x}} = \text{diag}(\mathbf{x}) (\boldsymbol{\lambda} + \mathbf{A} \mathbf{x}^{\mathbf{B}}) \quad (14)$$

with \mathbf{A} , \mathbf{B} , and $\boldsymbol{\lambda}$ of size $n \times m$, $m \times n$, and $n \times 1$. Just like before, the stoichiometric information is contained in one matrix and the kinetics in another. There is also a famous particular case of this kind of system, for $\mathbf{B} = \mathbf{I}$, Equation (14) becomes the Lotka–Volterra model for n species.

An important property of GLV systems is their invariance when subject to quasimonomial transformations $\mathbf{x} = \mathbf{y}^{\mathbf{C}}$, where \mathbf{C} is a square non-singular matrix. The result of this transformation is a GLV system itself:

$$\dot{\mathbf{y}} = \text{diag}(\mathbf{y}) (\hat{\boldsymbol{\lambda}} + \hat{\mathbf{A}} \mathbf{y}^{\hat{\mathbf{B}}}) \quad (15)$$

where

$$\begin{aligned} \hat{\mathbf{A}} &= \mathbf{C}^{-1} \mathbf{A} \\ \hat{\boldsymbol{\lambda}} &= \mathbf{C}^{-1} \boldsymbol{\lambda} \\ \hat{\mathbf{B}} &= \mathbf{B} \mathbf{C} \end{aligned} \quad (16)$$

All the systems that can be converted into one another through a quasimonomial transformation form a class of equivalence, sharing a great deal of important properties such as number of steady states and their stability (Hernández-Bermejo and Fairén, 1995).

A very complete account of the properties of this formalism can be found in Hernández-Bermejo et al. (1998), but we will describe two applications of the quasimonomial transformation, that are specially relevant in this context.

When matrix \mathbf{B} does not have full rank $r < n$, a transformation matrix can be chosen

$$\mathbf{C} = \begin{pmatrix} \mathbf{I}_{r \times r} & \boldsymbol{\phi}_1 & \dots & \boldsymbol{\phi}_k \\ \mathbf{0}_{n-r \times r} & & & \end{pmatrix} \quad (17)$$

where ϕ_i $i = 1 \dots k$ are basis vectors for the kernel of \mathbf{B} . The transformed exponents will in this case be $\hat{\mathbf{B}} = [\mathbf{B}_{m \times r} | \mathbf{0}_{m \times k}]$. From the structure of Equation (14), it follows that a number of variables in the transformed system equal to the dimension of $\ker(\mathbf{B})$ have no influence on any equation other than their own. These variables result in quadratures and can therefore be taken out of the system.

Any GLV can be converted to a Lotka–Volterra as a special case of quasi-monomial transformation $\mathbf{q} = \mathbf{x}^{\mathbf{B}}$, that results in the Lotka–Volterra model where the variables are replaced by the quasi-monomial terms:

$$\dot{\mathbf{q}} = \text{diag}(\mathbf{q}) (\mathbf{B}\lambda + \mathbf{B}\mathbf{A}\mathbf{q}) \quad (18)$$

Since the number of quasi-monomial is often greater than that of variables, matrix \mathbf{B} is seldom square and $\mathbf{B}\mathbf{A}$ will often be singular. From Equation (16) follows that any two systems from a class of equivalence will result in the same Lotka–Volterra representation, which can be taken to be a canonical form for the whole class. In the Lotka–Volterra systems, all non-linearities of the system are reduced to quadratic terms and any interaction term between two variables has the form $c q_i q_j$ where the constant c is the (i,j) -th entry of $\mathbf{B}\mathbf{A}$.

2.2.3. Relation between the Formalisms

Any system written as a power-law can be translated to a quasi-polynomial system and vice versa. This is trivial for small systems and can be done applying a formula to the matrices of arbitrarily large and complex systems (Marin-Sanguino et al., 2010). This similarity leads to many common properties that have been found using completely different methods in both formalisms. For instance, the symmetry matrix of an autonomous GMA or s-system (Voit, 1992) is the \mathbf{B} matrix of the corresponding GLV. The rank deficiency in such matrix, implies existence of parameter transformation groups that can decouple a power-law system the same way transformation (Equation 17) does with a GLV. From now on, we will consider both formalisms to be equivalent (Voit and Savageau, 1986), so we can talk, for instance, about the \mathbf{B} matrix of a GMA or the class of equivalence to which an s-system belongs.

2.3. Numerical Simulations

To verify the theoretical considerations, we simulated different non-linear models in s-system representation that were taken from the literature (Voit, 2000).

2.3.1. Integration of the Differential Equations with Perturbed Initial Values

Differential equations were numerically integrated with Matlab's ode15s solver. Integration time was estimated from the the biggest real eigenvalue of the Jacobian of the linearized, full system at its steady state:

$$t_{\text{end}} = -(\max(\text{Re}(\underline{\lambda})))^{-1} \cdot 5$$

The resulting trajectories of the slow variables—those not in quasi-steady-state (qss)—of the original system were compared

to the trajectories of the reduced system, in which the fast variables are assumed to be in qss. The robustness of the approximation was tested by performing simulations of the full system in which the qss variables had random initial values distant to qss by a factor of 10.

The value for the absolute perturbation of the fast variables is defined as the Euclidean norm of the natural logarithm of the quotients of the initial values of the fast variables $x_{f,0}$ and the corresponding quasi-steady state values at time 0, $x_{f,qss}$:

$$\delta y = \left\| \left(\ln \left(\frac{x_{f_1,0}}{x_{f_1,qss}} \right) \ln \left(\frac{x_{f_2,0}}{x_{f_2,qss}} \right) \dots \ln \left(\frac{x_{f_n,0}}{x_{f_n,qss}} \right) \right) \right\|$$

with n fast variables x_{f_i} .

2.3.2. Effect of the Perturbation of Fast Variables on the Slow Trajectories

In a next step, the effect of this perturbation was tested. The data of the slow trajectories was sampled at defined times for the full and the reduced system and the mean and standard deviation at these points in time were calculated for 1000 simulations with randomized initial values.

Additionally the trajectory of the slow variables in the original, full system was also interpolated and sampled at the same points in time. To get an objective measure of the relative error between reduced and full system, first the relative error of the trajectories in the reduced system $x(t)_{s_i,qssa}$ compared to the full system $x(t)_{s_i}$ was calculated for each of the slow variables by:

$$E_i = \frac{\int_0^{t_{\text{end}}} |x(t)_{s_i,qssa} - x(t)_{s_i}| dt}{\int_0^{t_{\text{end}}} |x(t)_{s_i}| dt}$$

with $x_{s_i,ss}$ being the steady-state-values that serve as a baseline for the comparison. The integral was numerically computed with the trapezoidal method, given the data from the trajectories. To get a number for the system considering all variables, the Euclidean norm of all these errors $|E|$ was plotted against δy . Each of the points represents one of the 1000 simulations.

2.4. Random Network Generation

In order to benchmark the methods for large systems, synthetic genetic networks were generated. When modeled as s-systems, this networks consists of a matrix of kinetic orders and a vector of turnover numbers as indicated below in the results section. The models were generated in python using the standard libraries scipy and numpy. The turnover numbers were generated at random in three groups to ensure the existence of three different time-scales. Each group was generated following a normal distribution with different means and standard deviations calculated from the distances between the means to guarantee the existence of three distinct. The number of variables in each group (time-scale) was also predetermined. The kinetic order matrices were generated as sparse matrices of density 0.05. Each network was tested to ensure stability and that all the components were connected (using the library Network X).

3. RESULTS

3.1. Log-Modes

The Jacobian of a system under a particular set of transformations (like the quasimonomial transformation) will always have the same eigenvalues as those of the original system. In the case of s-systems under the logarithmic transformation, the Jacobians are identical. S-systems can be explicitly rewritten after undergoing the transformation $\mathbf{y} = \log(\mathbf{x})$, the transformed equations take the form (Savageau, 1976),

$$\dot{\mathbf{y}} = \text{diag}(\alpha) \exp((G - I)\mathbf{y}) - \text{diag}(\beta) \exp((H - I)\mathbf{y})$$

The Jacobian matrix of the system defined in terms of \mathbf{y} at a steady state is identical to the Jacobian of the original system, so the coefficients defined in Equation (6) can be used to define a new set of modes that we will call log-modes (ℓ).

$$\log(\ell) = \mathbf{M}\mathbf{y}$$

which will actually be a monomial transformation:

$$\ell = \mathbf{x}^{\mathbf{M}}$$

So any fast or slow manifolds identified from the log modes will take the form of a power-law and can be back-substituted into any of the formalisms here discussed without generating algebraic constraints.

3.2. Identifying Time-Scales for the Variables through the S-System Representation

The existence of analytic steady-state solutions in s-systems makes it possible to apply the quasi-steady-state hypothesis to obtain the behavior of the slow part of multi-level systems (Savageau, 1976; Savageau and Sorribas, 1989), a very similar procedure has been used in the context of sensitivity analysis (Delgado and Liao, 1995). In this section we will generalize the procedure to split a dynamic system into its time scales, obtaining equations for the all of them without generating algebraic constraints. We will start using the s-system representation and will then move on to more general considerations.

Without loss of generality, the variables in Equation (12) can be arranged according to their f -factor in decreasing order, the variables can be classified as slow or fast by finding a variable x_k such that $\|f_{k+1} - f_k\|$ is maximal. Now a non-dimensionalization for time can be applied $\tau = f_k t$

$$\frac{dz_i}{d\tau} = \frac{f_i}{f_k} \left(\prod_j z_j^{g_{ij}} - \prod_j z_j^{h_{ij}} \right)$$

defining $\varepsilon = \frac{f_{k+1}}{f_k}$, the multiplier for the first k equations becomes: $\frac{f_1}{f_k} > \frac{f_2}{f_k} > \dots > 1$ and the rest $\varepsilon > \varepsilon \frac{f_{k+2}}{f_{k+1}} > \dots > \varepsilon \frac{f_n}{f_{k+1}}$.

$$\begin{aligned} \frac{dz_i}{d\tau} &= \hat{f}_i \left(\prod_j z_j^{g_{ij}} - \prod_j z_j^{h_{ij}} \right) & i = 1, \dots, k \\ \frac{dz_i}{d\tau} &= \varepsilon \hat{f}_i \left(\prod_j z_j^{g_{ij}} - \prod_j z_j^{h_{ij}} \right) & i = k + 1, \dots, n \end{aligned} \quad (19)$$

if $f_k \gg f_i \Rightarrow \frac{dz_i}{d\tau} = \varepsilon (\mathbf{z}^{\mathbf{G}} - \mathbf{z}^{\mathbf{H}})$, which enables to get a quasi-steady-state (qss) solution for the fast variable. s-systems share the telescopic property discussed above for linear systems so the algebraic constraints generated by the qss assumption can be back-substituted in the system as shown in the Appendix (Supplementary Material). As a result, the system can be divided in two, a fast system:

$$\frac{dz_i}{d\tau} = \hat{\alpha}_i \prod_{j=1}^k z_j^{g_{ij}} - \hat{\beta}_i \prod_{j=1}^k z_j^{h_{ij}} \quad i = 1, \dots, k \quad (20)$$

where the slow variables are taken as constants and grouped into $\hat{\alpha}_i$ and $\hat{\beta}_i$. The normalized steady state is no longer at one, since it depends on the values assigned to the slow variables.

A time rescaling $T = \varepsilon \tau$ provides the complementary time scale. The slow system that depends exclusively on the slow variables:

$$\frac{dz_i}{dT} = \hat{f}_i \left(\prod_{j=k}^n z_j^{g_{ij}} - \prod_{j=k}^n z_j^{h_{ij}} \right) \quad i = k + 1, \dots, n \quad (21)$$

See Supplementary Information for a detailed calculation.

This procedure can only be applied to an s-system but it provides information of use for the more general cases. GLV systems with a single equilibrium point can be exactly rewritten as s-systems (Hernández-Bermejo and Fairén, 1995), s-systems can also dominate the dynamics of arbitrary non-linear systems in a well defined region of their parameter space (Savageau et al., 2009) or arise as good approximations through a Taylor series (Savageau, 1969a). The validity of the turnover numbers as indicators for timescales is in fact so robust, that the inverse of the turnover, the transition time, was defined as a reference in the model free setting of biochemical enzyme assays (Easterby, 1981). Turnover numbers are only a valid approach for well behaved systems in which they dominate over the rest of the equations, the next section will deal with not so well behaved systems.

3.3. Collinearity among the Quasinomials

In order to assess whether a system is “well behaved” in the sense mentioned above, a closer examination of \mathbf{B} is in order. Sensitivity to parameter combinations can be assessed through the spectrum of the corresponding matrix (Hearne, 1985).

Since the matrices involved are not usually square, the SVD decomposition of the matrix, $\mathbf{B} = \mathbf{U}_B \mathbf{\Sigma}_B \mathbf{V}_B^T$, will be of great use.

It has been seen that a rank deficiency in \mathbf{B} allows to decouple some of the variables of the system. This results in an invariant manifold spanned by the corresponding vectors of \mathbf{V}_B which can consist of infinite equilibrium points or preclude any sort of equilibrium (Voit, 1991). When the matrix is not singular but it is ill conditioned, a similar phenomenon happens. This can be seen by applying a quasimonomial transformation:

$$\mathbf{x} = \mathbf{y}^{\mathbf{V}_B} \quad (22)$$

where each of the new variables y_i is associated to a singular vector. The exact dynamics of all these new variables will have GLV form as shown in Equation (16). From inspection of the resulting system

$$\dot{\mathbf{y}} = \text{diag}(\mathbf{y}) \left(\hat{\boldsymbol{\lambda}} + \hat{\mathbf{A}} \mathbf{y}^{\mathbf{U}_B \mathbf{\Sigma}_B} \right) \quad (23)$$

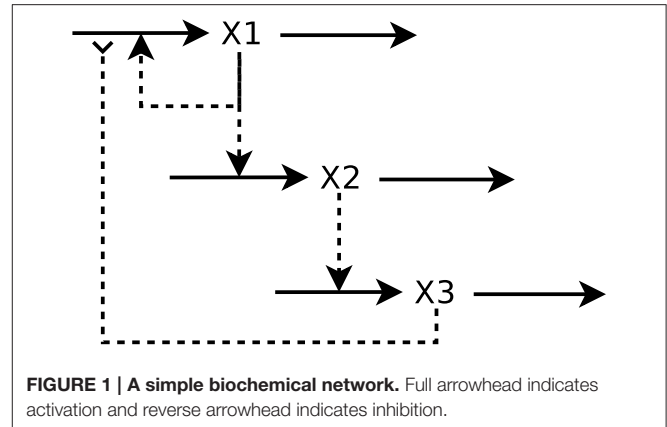
it is straightforward to see that the exponents of y_i in all monomial terms are multiplied by σ_i , as the later tends to zero, the variable will lose influence on the all the other variables, reducing the real dimension of the system.

3.4. The Stoichiometric and Kinetic Components of the Invariant Matrix

Analyzing the log-modes of a non-linear system at a certain equilibrium point has the risk of not being representative, since its Jacobian may change dramatically when it moves away from the linear region. As has been seen above, the turnovers of the variables and the singular vectors of \mathbf{B} provide two complementary methods. The interplay between these three alternative representations can be seen in the LV of the corresponding equivalence class. The constant matrix $\mathbf{B} \mathbf{A}$ does not result from a linearization, it defines all interactions between variables the whole phase space. Although there is no closed form for the singular/eigen-values of a matrix product, a great deal can be learned by calculating the SVD of both \mathbf{A} and \mathbf{B} :

$$\mathbf{B} \mathbf{A} = \mathbf{U}_B \mathbf{\Sigma}_B \mathbf{W} \mathbf{\Sigma}_A \mathbf{V}_A^T \quad (24)$$

where $\mathbf{W} = \mathbf{V}_B^T \mathbf{U}_A$. The three matrices \mathbf{U}_B , \mathbf{W} , and \mathbf{U}_A are unitary and will not amplify or dampen any perturbation to the variables. Any change in the norm of the perturbation will come from the two diagonal matrices of singular values, one coming from the stoichiometric component of the system, $\mathbf{\Sigma}_A$, and one from the kinetic $\mathbf{\Sigma}_B$. When only one of these matrices has extreme values it will dominate the response of the system and one of the two methods mentioned above will be accurate. No sudden changes of the jacobian are to be expected, since the Jacobian of an normalized LV system is precisely (Equation 24), see Diaz-Sierra et al. (1999). When both sets of singular values are in the same range, the system will not be decomposable by time hierarchies. Extreme cases, where both sets of singular values have big differences, will result in systems where the time hierarchies shift along the orbits of the system. In such cases,



Equation (24) would be a good starting point to identify regions of interest in the parameter and in the phase space.

3.5. A Simple Example

Lets start with a model of a small regulatory network of three genes that affect one another's induction as depicted in **Figure 1**. Obtaining the GMA model is straightforward:

$$\begin{aligned} \dot{x}_1 &= \alpha_1 x_1^{g_{1,1}} x_2^{-g_{1,3}} - \beta_1 x_1 \\ \dot{x}_2 &= \alpha_2 x_1^{g_{2,1}} - \beta_2 x_2 \\ \dot{x}_3 &= \alpha_3 x_2^{g_{3,2}} - \beta_3 x_3 \end{aligned} \quad (25)$$

which can be rewritten as a GLV by just factoring the variables out:

$$\begin{aligned} \dot{x}_1 &= x_1 \left(\alpha_1 x_1^{g_{1,1}-1} x_2^{-g_{1,3}} - \beta_1 \right) \\ \dot{x}_2 &= x_2 \left(\alpha_2 x_1^{g_{2,1}} x_2^{-1} - \beta_2 \right) \\ \dot{x}_3 &= x_3 \left(\alpha_3 x_2^{g_{3,2}} x_3^{-1} - \beta_3 \right) \end{aligned} \quad (26)$$

So

$$\mathbf{B} = \begin{pmatrix} g_{11} - 1 & 0 & -g_{13} \\ g_{21} & -1 & 0 \\ 0 & g_{32} & -1 \end{pmatrix} \quad (27)$$

Computing the turnovers of the variables in Equation (25) is straightforward:

$$f_i = \beta_i \quad \forall i$$

normalizing:

$$\begin{aligned} \dot{z}_1 &= \beta_1 (z_1^{g_{1,1}} z_2^{-g_{1,3}} - z_1) \\ \dot{z}_2 &= \beta_2 (z_1^{g_{2,1}} - z_2) \\ \dot{z}_3 &= \beta_3 (z_2^{g_{3,2}} - z_3) \end{aligned} \quad (28)$$

The Jacobian matrix is:

$$\mathbf{J} = \text{diag}(\boldsymbol{\beta}) \mathbf{B}$$

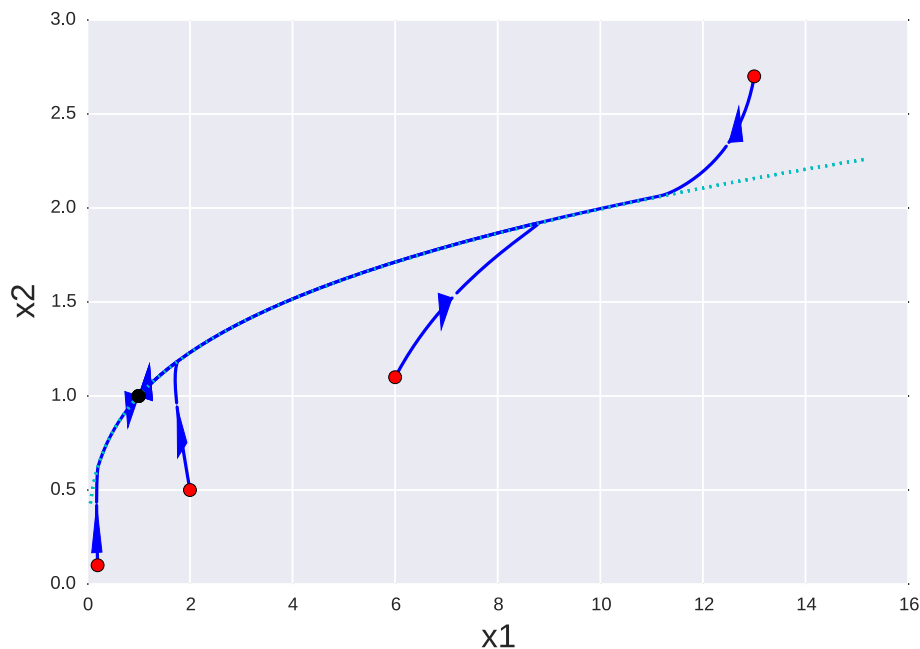


FIGURE 2 | Orbits of a simple genetic network for different initial conditions (red dots) projected onto the $x_1 - x_2$ plane. Orbits clearly evolve toward a one dimensional manifold (dotted line) that follows to the singular vector v_3 corresponding to the smallest singular value. The equilibrium point $\mathbf{x} = \mathbf{1}$ is marked as a black dot. Red dots mark starting points of different orbits and the black dot is the equilibrium point they all tend to.

for the particular case $\beta_i = 1 \forall \beta$ all the turnover numbers are also 1 and all variables are expected to operate in the same time scale. However, as can be seen in **Figure 2**, where a special case is simulated— $g_{11} = 1.1$, $g_{13} = 0.48$, $g_{21} = 0.3$, $g_{32} = 0.7$ —the system approaches a slow manifold defined by the singular vector with the smallest singular value of \mathbf{B} : $\mathbf{v} = (0.939, 0.281, 0.197)$. So the system has a slow manifold. Transformation using Equation (22), an alternative formulation is obtained with variables (y_1, y_2, y_3) and matrices:

$$\lambda = \begin{pmatrix} -0.050 \\ 0.99 \\ -1.4 \end{pmatrix} \quad (29)$$

$$\mathbf{A} = \begin{pmatrix} -0.10 & 0.78 & -0.62 \\ 0.33 & -0.57 & -0.76 \\ 0.94 & 0.28 & 0.20 \end{pmatrix} \quad (30)$$

$$\mathbf{B} = \begin{pmatrix} 0.29 & 0.40 & -0.00056 \\ -0.81 & 0.66 & 0.00019 \\ 1.2 & 0.36 & 0.00027 \end{pmatrix} \quad (31)$$

and as can be seen by the small exponents of y_3 , this variable has negligible influence on the dynamics of the other two, **Figures 3, 4**.

The same procedure can be done applying the decomposition shown in Equation (3.1) to obtain the equations for the log modes. In this case, the results are very similar to those already shown, since the Jacobian matrix of the system $\mathbf{J} = \mathbf{B}$. Even though the similarity decomposition that defined the log-modes is not equal to svd decomposition, the coefficients of the slowest log-mode of the system are within 0.5% of those of v_3 . Additional

simulations show that special cases with well conditioned \mathbf{B} led to similar time-scales using turnover numbers and log-modes, for cases with similar turnovers, the log-modes are similar to the slow manifolds predicted by \mathbf{B} , as can be expected from Equation (24)—data not shown.

3.6. A Large Network

The toy model shown above is useful to understand the theory behind the methods, but in order to test the performance of the method on large scale models, randomly generated genetic networks were used. Genetic networks can be modeled using s-systems of the form $\dot{x}_i = \alpha_i \prod_j x_j^{g_{ij}} - \beta_i x_i$, where the interactions between genes are concentrated in the kinetic orders of the positive term. The turnover can be factored out:

$$\dot{x}_i = F_i \left(\prod_j x_j^{g_{ij}} - x_i \right) \quad (32)$$

The details of how the networks were generated are shown in the methods section, and the results were satisfactory in all cases. Here we will show the analysis of a representative network with 75 variables divided in three time scales with turnover numbers of 1, 100, and 10^4 . The number of variables in each group (time-scale) was 10, 25, and 40 respectively. The network is defined by 356 parameters: 281 non-zero kinetic orders and 75 turnover numbers. The parameters values of the s-system model are provided as supplementary data.

The existence of three time scales opens several possibilities. If the model is to be partitioned in two, the variables

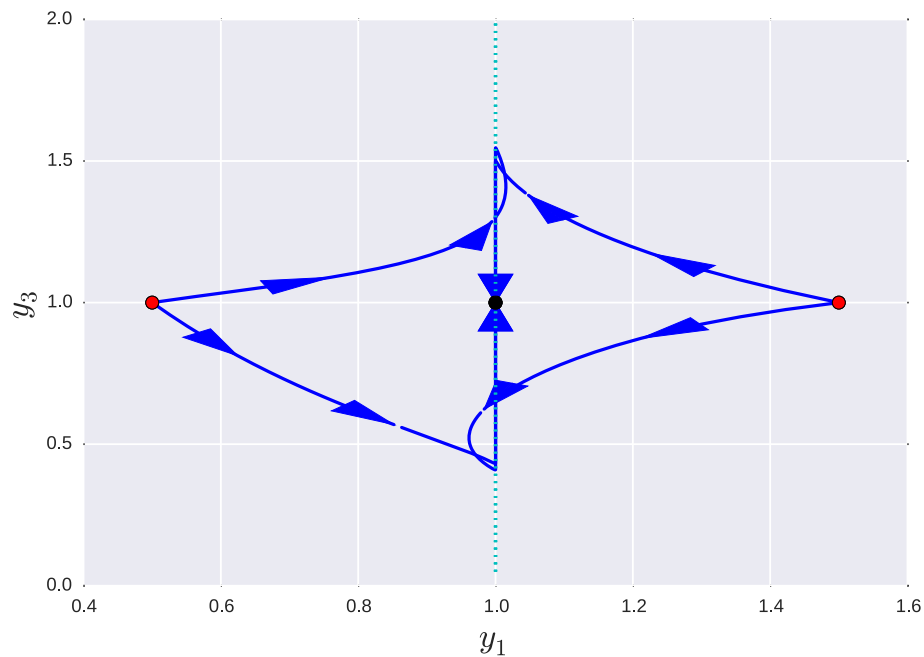


FIGURE 3 | Projection of the orbits of a the transformed system onto the $y_1 - y_3$ plane for different initial conditions. In the new coordinate system $x = y^V$, each axis correspond to a singular value, so the manifold shown in the previous figure, is now parallel to the y_3 axis (dotted line). Red dots mark starting points of different orbits and the black dot is the equilibrium point they all tend to.

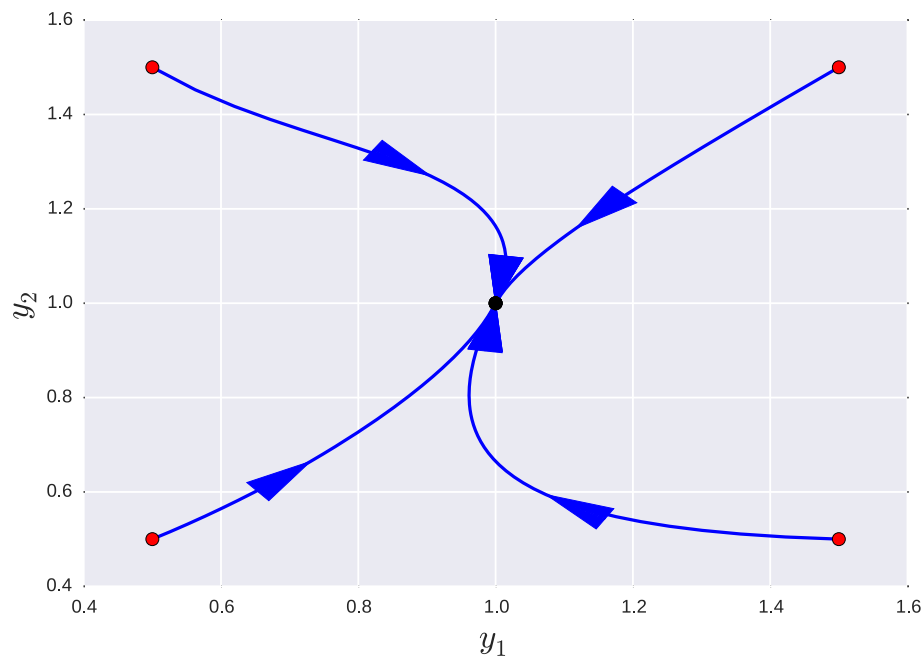
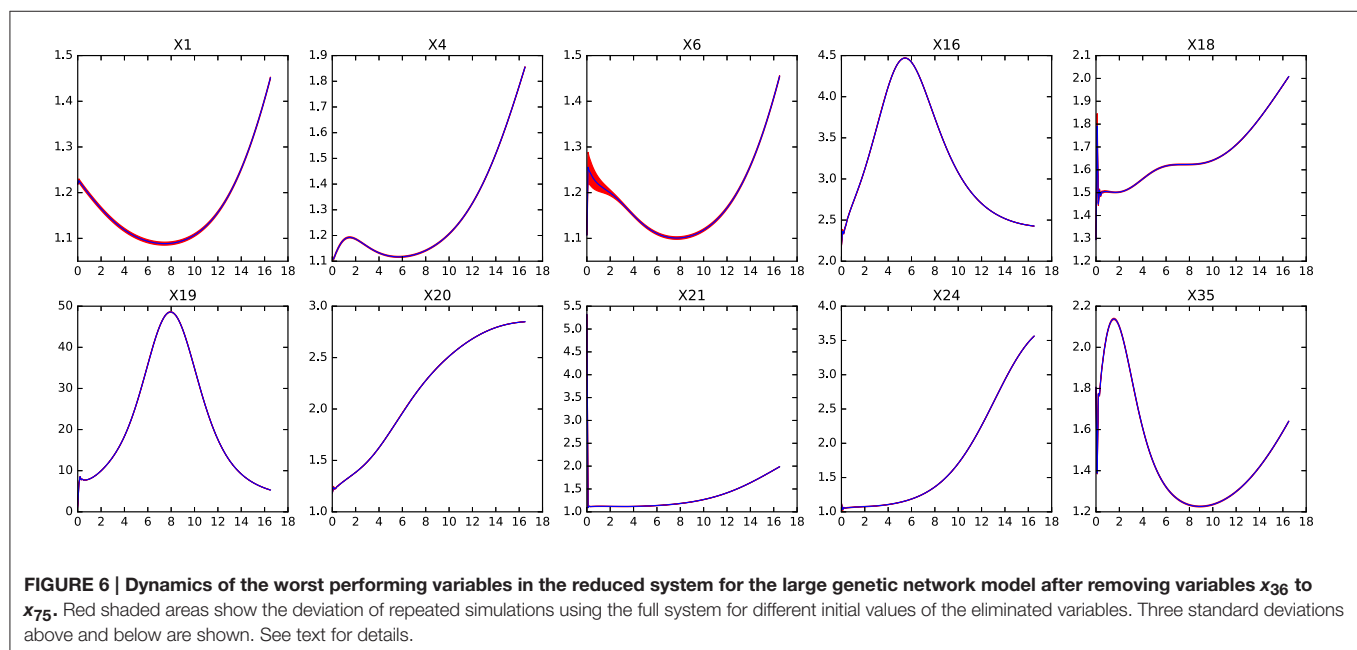
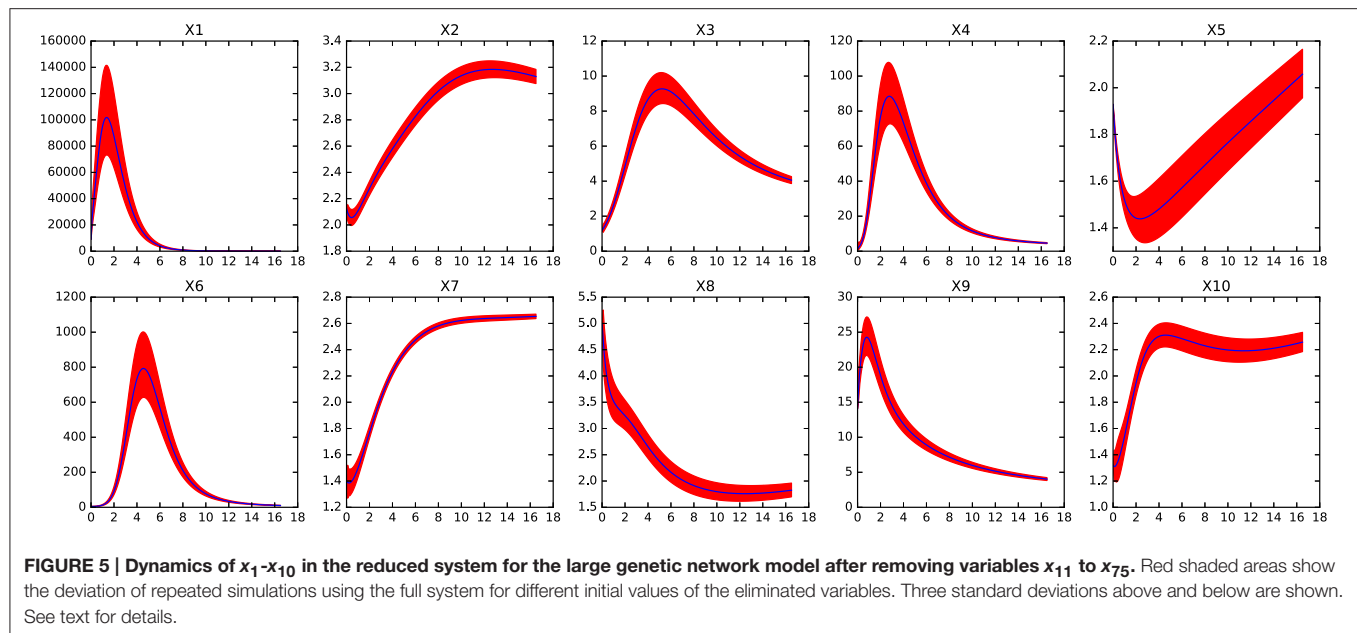


FIGURE 4 | Projection of the orbits of a the transformed system onto the $y_1 - y_2$ for different initial conditions. In this projection, the manifold appears as a point aligned with the equilibrium point. Changing the values of y_3 within an order of magnitude resulted in no appreciable change in this projection, since the third variable is *de facto* uncoupled. Red dots mark starting points of different orbits and the black dot is the equilibrium point they all tend to.

in the middle range can be assigned to the fast or slow subsystem. Moreover, successive separation can lead to three different submodels, one per time scale. Each approach will

generate systems with different accuracy and degree of stiffness, so the optimal decision will depend on the goal of the analysis.



Figures 5, 6 show the errors in the dynamics of 100 different simulations of the two possible slow systems. In the first case, a system with only ten variables is obtained, in the second, the final number of slow variables is thirty five. Figure 7 shows the accumulated error. As can be seen, the smallest model has a much higher error but still agrees qualitatively with the dynamics of the full system. The bigger model, has an extremely small error but it still contains variables operating in two different time scales. This increases the computational cost of integration as shown in Table 1. The bigger model provides high accuracy with a substantial improvement in computational cost and a significant

reduction in complexity and the number of variables. Since most biological measurements are subjected to high levels of noise, the smallest and much simpler model system will often be adequate as well.

Finally, the network can be split into three different submodels able to reproduce the slow, middle and fast dynamics respectively. Figure 8 shows how the reduction processes affects the connectivity of the network. Submodels of the fast dynamics, reduce the degree of connectivity, since many connections between fast variables happen through slow variables that are frozen in the fast time scales.

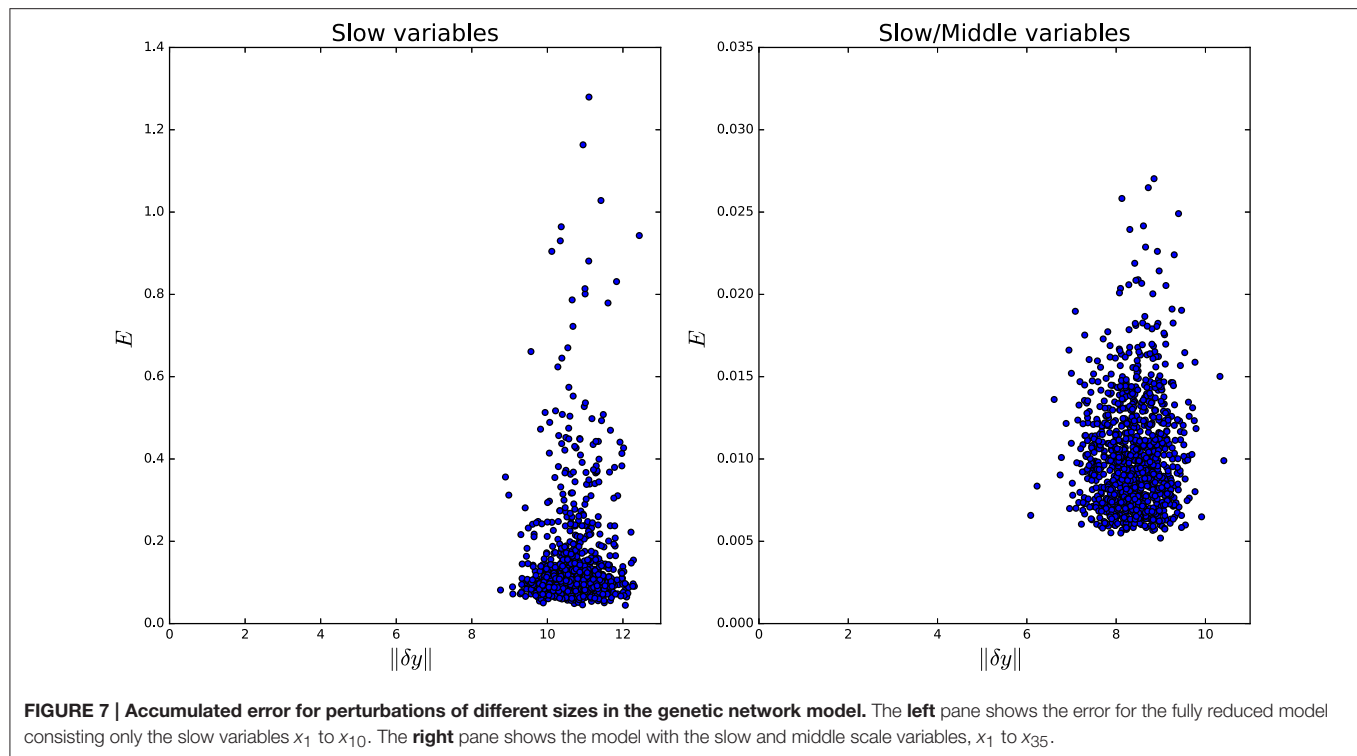


TABLE 1 | Comparison of simulation times between original and reduced models of the large randomized networks.

| Genetic network | | |
|-----------------|----------------|----------------------|
| | Time | QSS |
| Full | 1109 ± 400 | none |
| Slow | 3.6 ± 0.2 | x_{11} to x_{75} |
| Slow/Middle | 193 ± 6 | x_{36} to x_{75} |

Submodels of slower timescales, experience the opposite effect, since the variables that are eliminated through the quasi-steady-state assumption become links between slow variables.

3.7. Examples from Real Models

In order to test the applicability to real cases, three classical s-system models from the literature (Voit, 2000) were taken as examples for benchmarking: A very simplified model for the anaerobic fermentation of *Saccharomyces cerevisiae* with 5 variables, a model for the purin metabolism in man consisting of 16 variables and one for the tricarboxylic acid cycle in *Dictyostelium discoideum* constituted by 13 variables. These three models have been also used for benchmarking an alternative method of model reduction, which will enable further comparisons.

All three models had well conditioned **B** matrices, so timescales were assigned to each variable according to their turnover number.

3.7.1. Yeast

Eliminating the two fastest metabolites of this model of yeast glycolysis results in a robust reduced system that still can reproduce the slow dynamics with great accuracy (well within experimental error), as can be seen in **Figures 9, 10**. Even a perturbation δy of 3 still results in less than 14% error **E**.

3.7.2. TCA Cycle

The system is also reduced to less than two thirds of its size and results in good quantitative agreement with the full system. **Figure 11** shows how some variables reproduce the dynamic perfectly while x_6 and x_8 go through a short adaptation phase where their dynamics are not as robust as the rest. Accumulated error is shown in **Figure 12**.

3.7.3. Purine Metabolism

In this example, a more conservative approach is shown, where eliminating only a small set of the total number of variables shows a great quantitative agreement between the full and the reduced systems. **Figure 13** shows the only variable where an appreciable difference between the systems can be found. Accumulated error shown in **Figure 14**.

3.7.4. Performance

Two further metrics will be considered to evaluate the performance of the method: the reduction in simulation times due to the model reduction and the amount of variables eliminated in comparison to the alternative method (Liu et al., 2013) similar method. There is, to our knowledge only one method that has exploited the regular structure of

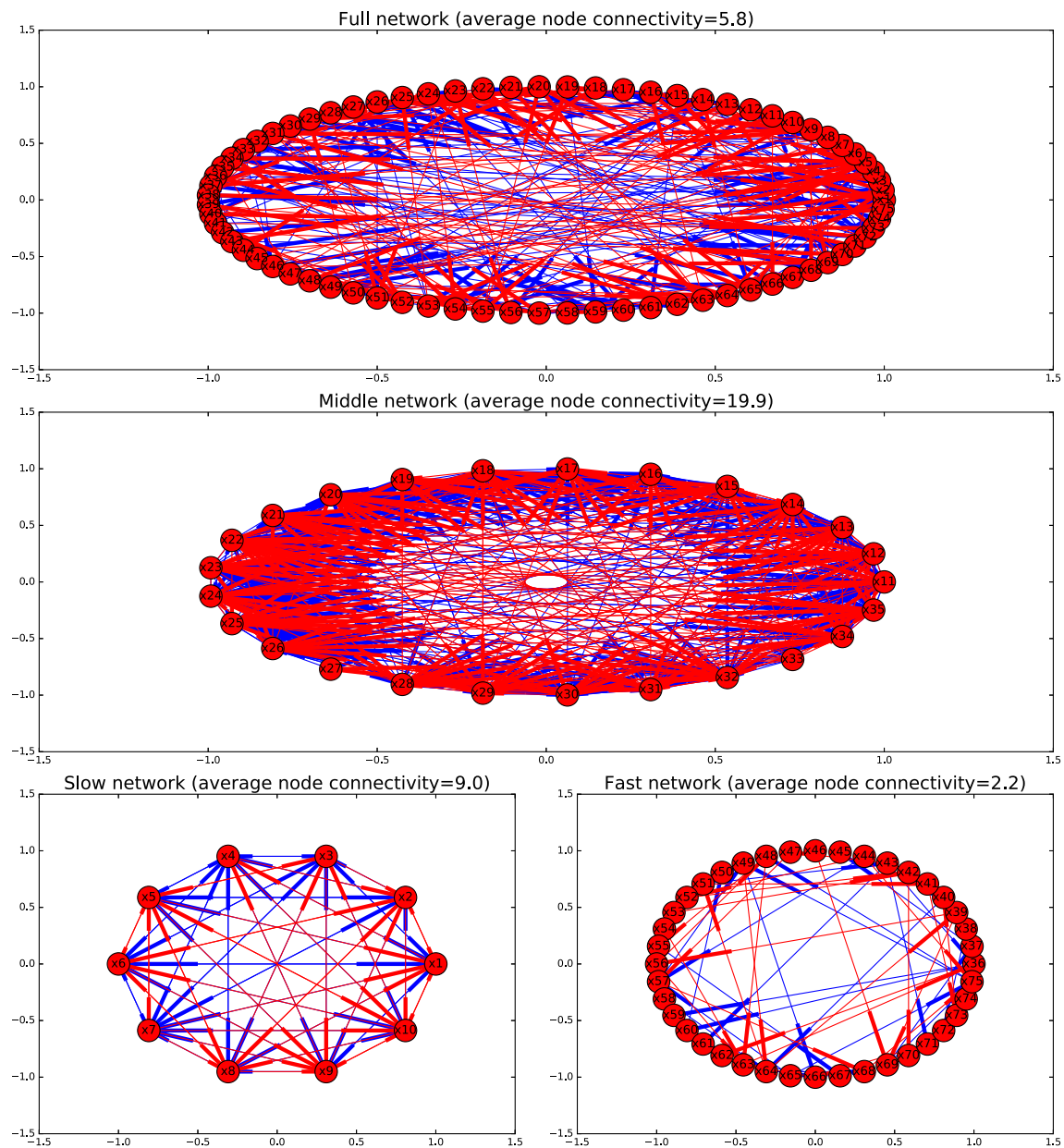


FIGURE 8 | Graphs depicting the different models as directed graphs. Activating interactions are shown in blue, inhibitory interactions in red.

canonical models to produce a model reduction algorithm (Liu et al., 2013). The alternative method does not provide a separation into submodels, it concerns itself exclusively with the elimination of variables using multicriteria optimization based on reactive weight, sensitivity, and flux analyses. Based on such optimization, the model is reformulated to eliminate one or more variables. For the sake of comparison, the methods presented in this study were used to obtain reduced models with total accumulated errors that were comparable to those of the previously mentioned approach, the number of variables that each method was able to remove is then compared.

Table 2 shows that model reduction always resulted in a significant improvement on the simulation times. Moreover, the number of reduced variables is always higher than or equal to the much more complex (and computationally demanding) existing method.

4. DISCUSSION

One of the bottlenecks for modeling biological systems is the need to find values for a great amount of parameters that cannot be measured directly. That, and the impossibility to predict how a change in the value of such parameters will change the dynamics

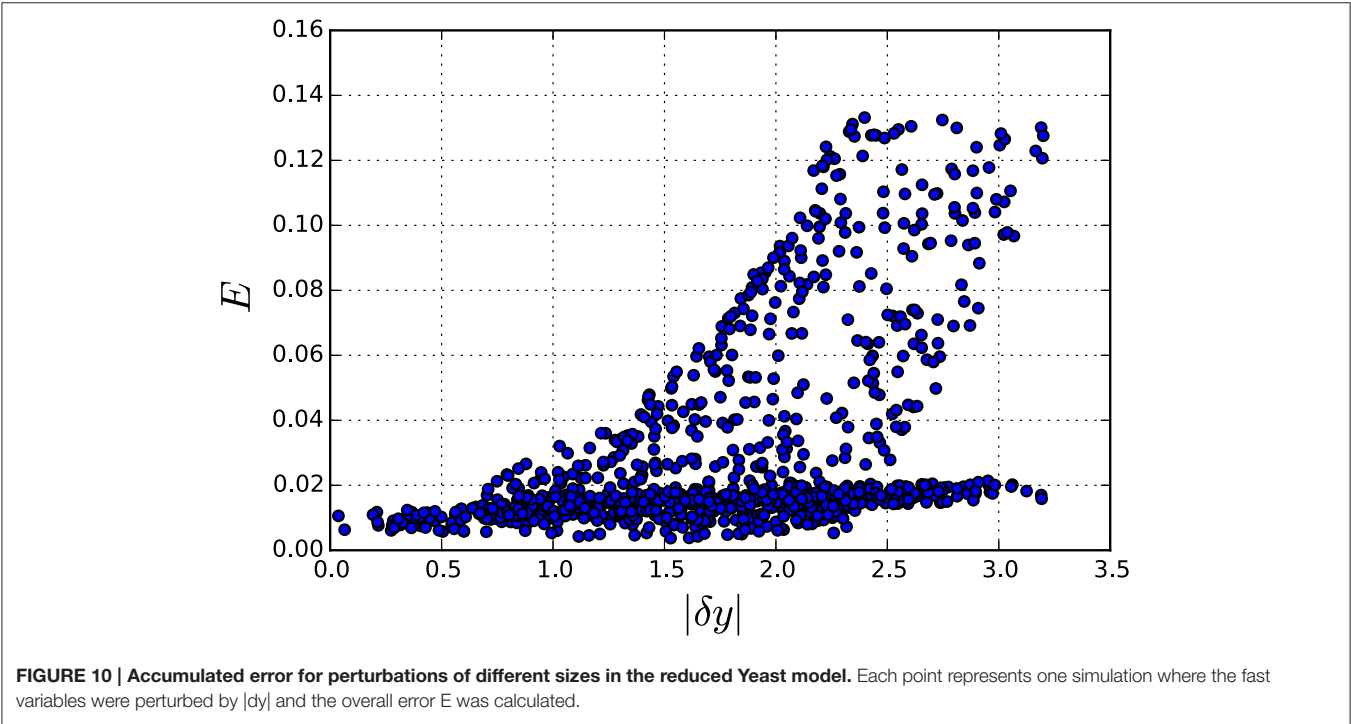
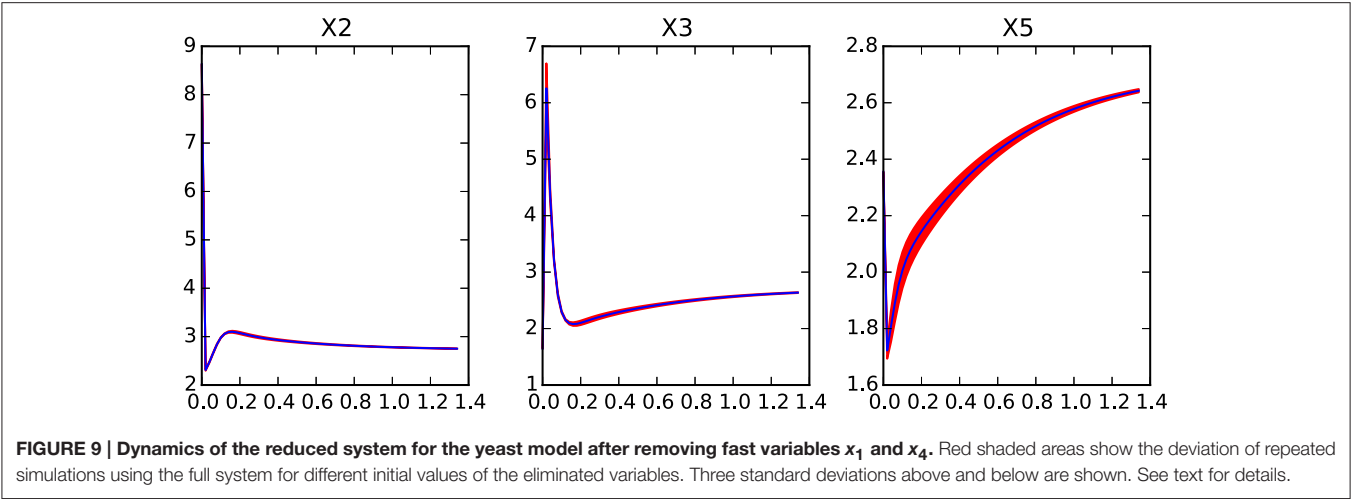
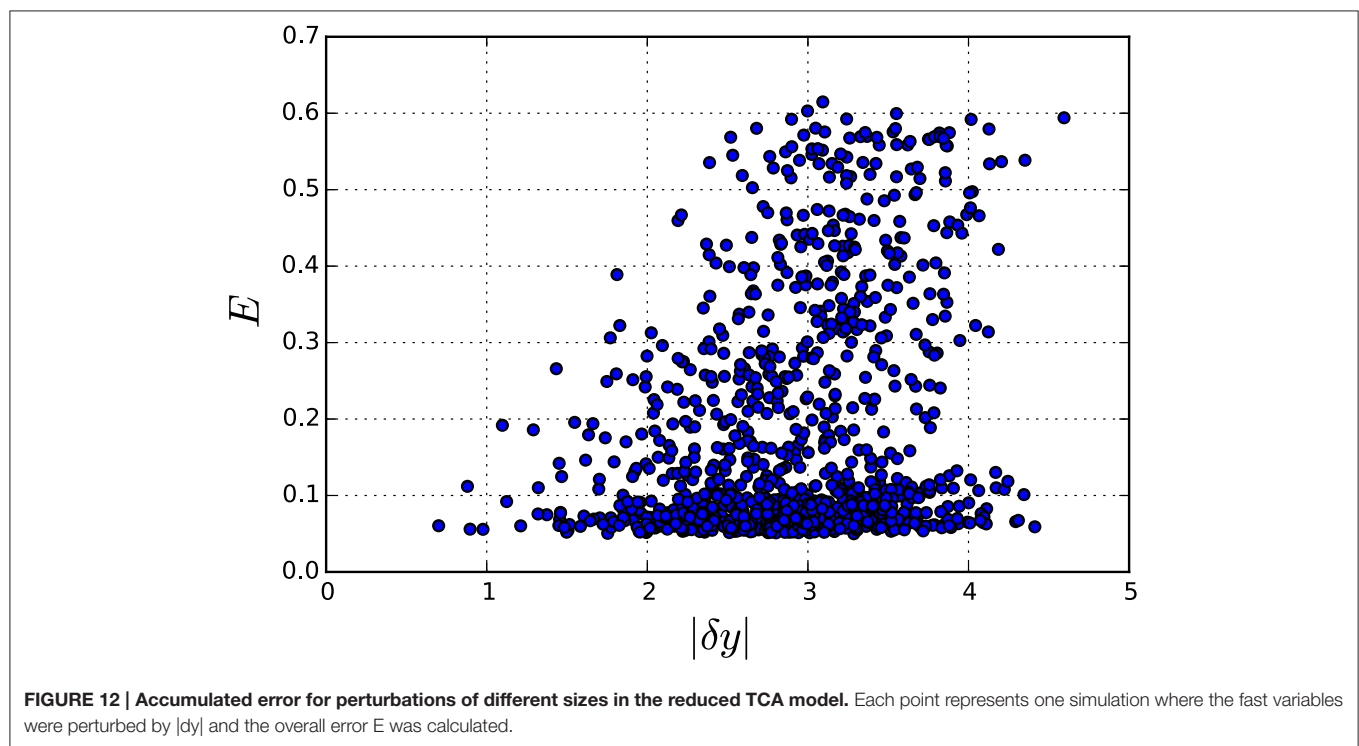
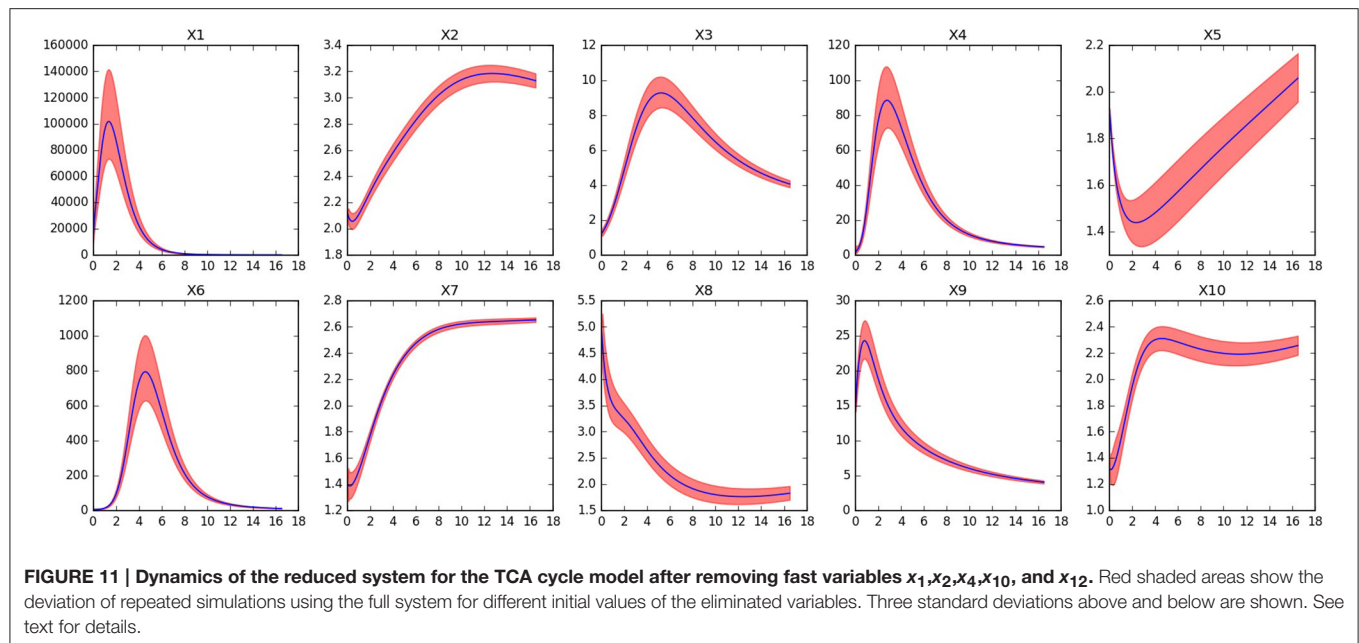


TABLE 2 | Comparison of simulation times between original and reduced models of the Yeast, TCA and purine models.

| | Simulation time | | Eliminated variables | |
|---------|-----------------|---------------|---------------------------------|-----------------------|
| | Full model | Reduced model | This study | Liu et al., 2013 |
| Yeast | 31.5 ± 1.0 | 11.3 ± 0.6 | x_1, x_2, x_4, x_5 | No reduction |
| TCA | 71.1 ± 7.4 | 50.9 ± 8.2 | $x_1, x_2, x_4, x_{10}, x_{12}$ | x_7 |
| Purines | 492.1 ± 46.6 | 260.8 ± 45.7 | x_3, x_8, x_{13} | x_6, x_{14}, x_{16} |

of the system, limit the reliability of numerical simulations. It is therefore imperative to find reliable tools for the global analysis and model reduction for non-linear systems.

Canonical, non-linear systems are flexible enough to reproduce any kind of non-linear behavior and, at the same time, all the information defining a particular model is encoded in two matrices and a vector. Methods like recasting (Savageau and Voit, 1987; Hernández-Bermejo et al., 1998) enable to rewrite virtually any non-linear system in one of the canonical forms treated here. Moreover, Design Space Analysis (Savageau et al., 2009) enables to decompose the parameter space of any system into qualitatively similar regions, each described by an s-system. These formalisms offer the exciting possibility of converting very abstract problems into simple linear algebra operations. Converting topologically equivalent systems into one another is done with three simple matrix products and identifying a slow manifold can be done by Singular Value Decomposition.



Moreover, any model in one of these formalisms can be exactly converted into a set of Lotka–Volterra equations. In the Lotka–Volterra representation, a single constant matrix determines the interactions between variables for the whole phase space, as opposed to a linearization, where the constant matrix is merely a local representation in the vicinity of an equilibrium. Decomposing this matrix into its kinetic and stoichiometric parts, provides a great deal of insight into the structure of the

system through the examination of the two corresponding sets of singular values. These results obtained with simple linear algebra, are as good as those that can be obtained using much more complicated approaches (Liu et al., 2013) as well as more general. The significance of this can best be appreciated in light of an example attributed to professor Grötschel (Holdren et al., 2010), a certain linear programming problem that would take 82 years to be solved by a computer in 1988 would be solved

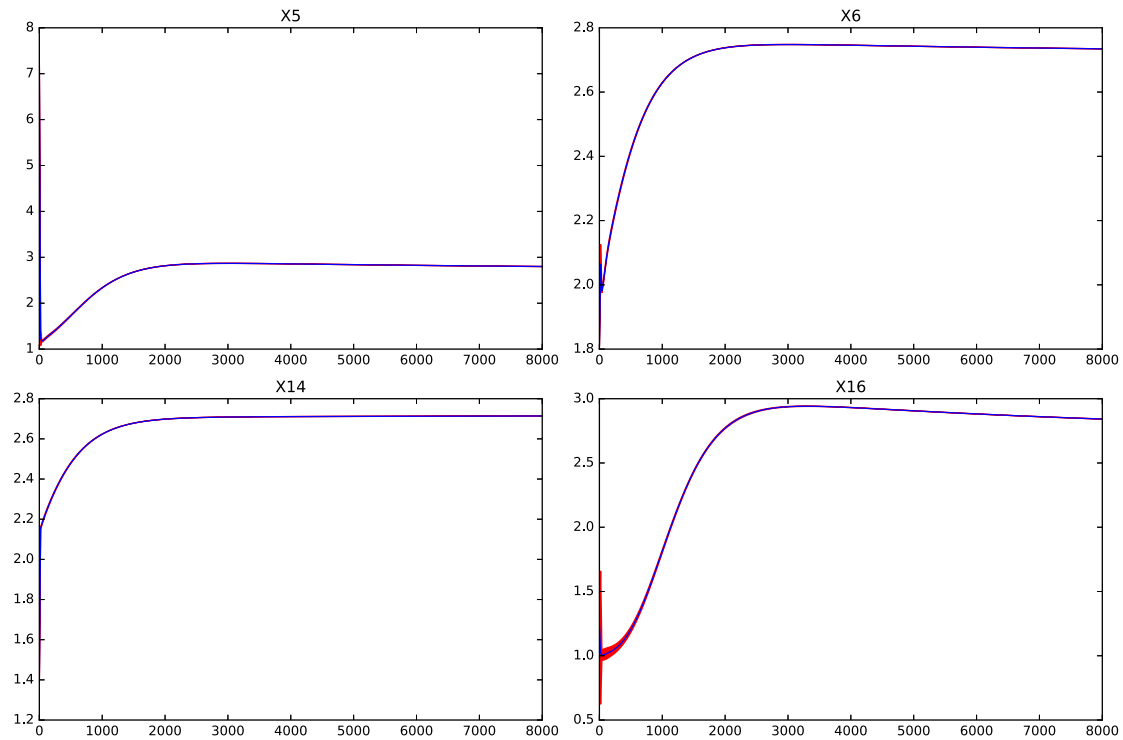


FIGURE 13 | Dynamics of x_{16} in the reduced system for the Purines metabolism model after removing fast variables x_3, x_8 , and x_{13} . Red shaded areas show the deviation of repeated simulations using the full system for different initial values of the eliminated variables. Three standard deviations above and below are shown. See text for details.

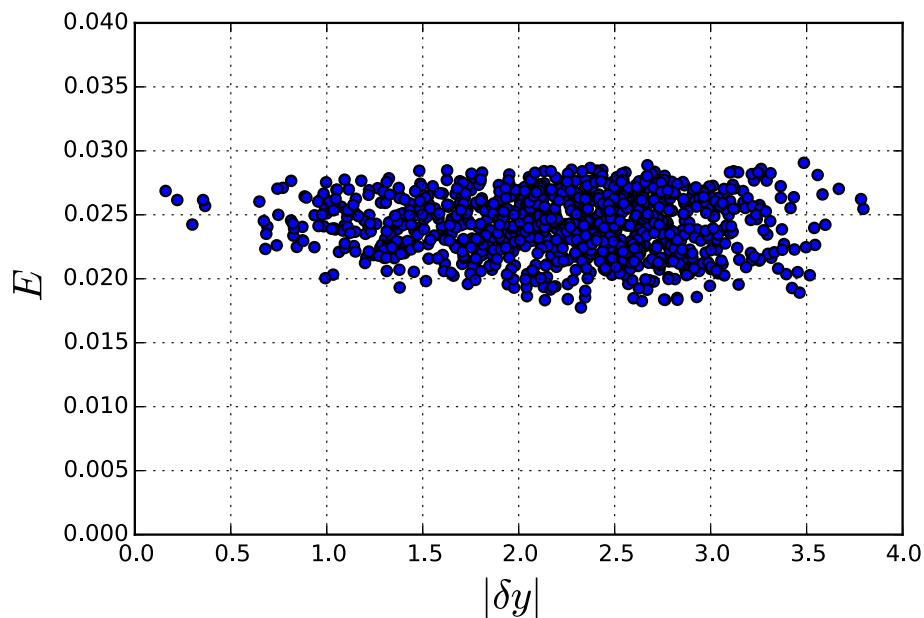


FIGURE 14 | Accumulated error for perturbations of different sizes in the reduced Purines metabolism model. Each point represents one simulation where the fast variables were perturbed by $|\delta y|$ and the overall error E was calculated.

in roughly a minute by a modern computer 15 years later. Of this improvement by a factor of 43 million, 1000 could be attributed to hardware improvements and the remaining 43,000 to improvements in numerical algorithms, mostly numerical linear algebra.

AUTHOR CONTRIBUTIONS

AK and AM proposed the topic and guided the research. HL and AM developed the concept and carried out the simulations and data analysis. All authors participated in the writing of the manuscript and approved the final version.

REFERENCES

- Delgado, J., and Liao, J. C. (1995). Control of metabolic pathways by time-scale separation. *Biosystems* 36, 55–70.
- Díaz-Sierra, R., Hernández-Bermejo, B., and Fairén, V. (1999). Graph-theoretic description of the interplay between non-linearity and connectivity in biological systems. *Math. Biosci.* 156, 229–253.
- Easterby, J. (1981). A generalized theory of the transition time for sequential enzyme reactions. *Biochem. J.* 199, 155.
- Gerdtsen, Z. P., Daoutidis, P., and Hu, W.-S. (2004). Non-linear reduction for kinetic models of metabolic reaction networks. *Metab. Eng.* 6, 140–154. doi: 10.1016/j.ymben.2003.11.003
- Hearne, J. (1985). Sensitivity analysis of parameter combinations. *Appl. Math. Model.* 9, 106–108.
- Hek, G. (2010). Geometric singular perturbation theory in biological practice. *J. Math. Biol.* 60, 347–386. doi: 10.1007/s00285-009-0266-7
- Hernández-Bermejo, B., and Fairén, V. (1995). Nonpolynomial vector fields under the Lotka–Volterra normal form. *Phys. Lett. A* 206, 31–37.
- Hernández-Bermejo, B., Fairén, V., and Brenig, L. (1998). Algebraic recasting of nonlinear systems of odes into universal formats. *J. Phys. A* 31, 2415–2430.
- Holdren, J., Lander, E., and Varmus, H. (2010). *Report to the President and Congress: Designing a Digital Future: Federally Funded Research and Development in Networking and Information Technology*. Available online at: <https://www.whitehouse.gov/>
- Jamshidi, N., and Palsson, B. Ø. (2008). Top-down analysis of temporal hierarchy in biochemical reaction networks. *PLoS Comput. Biol.* 4:e1000177. doi: 10.1371/journal.pcbi.1000177
- Liu, Y.-C., Lin, C.-L., and Chuang, C.-H. (2013). An approach for model reduction of biochemical networks. *Comput. Biol. J.* 2013:263973. doi: 10.1155/2013/263973
- Lomnitz, J. G., and Savageau, M. A. (2015). Elucidating the genotype–phenotype map by automatic enumeration and analysis of the phenotypic repertoire. *NPJ Syst. Biol. Appl.* 1, 15003. doi: 10.1038/npsba.2015.3
- Marin-Sanguino, A., Mendoza, E. R., and Voit, E. O. (2010). Flux duality in nonlinear gma systems: Implications for metabolic engineering. *J. Biotechnol.* 149, 166–172. doi: 10.1016/j.jbiotec.2009.12.009
- Müller, S., Feliu, E., Regensburger, G., Conradi, C., Shiu, A., and Dickenstein, A. (2016). Sign conditions for injectivity of generalized polynomial maps with applications to chemical reaction networks and real algebraic geometry. *Found. Comput. Math.* 16, 69–97. doi: 10.1007/s10208-014-9239-3
- Palsson, B., Joshi, A., and Ozturk, S. (1987). Reducing complexity in metabolic networks: making metabolic meshes manageable. *Fed. Proc.* 46, 2485–2489.

ACKNOWLEDGMENTS

AM acknowledges funding from the German Federal Ministry of Education and Research (BMBF), eBio initiative 0316197 as well as the TUM Junior Fellow Fund.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: <http://journal.frontiersin.org/article/10.3389/fgene.2016.00166>

- Savageau, M., and Sorribas, A. (1989). Constraints among molecular and systemic properties: implications for physiological genetics. *J. Theor. Biol.* 141, 93.
- Savageau, M. A. (1969a). Biochemical systems analysis: I. Some mathematical properties of the rate law for the component enzymatic reactions. *J. Theor. Biol.* 25, 365–369.
- Savageau, M. A. (1969b). Biochemical systems analysis: II. The steady-state solutions for an n-pool system using a power-law approximation. *J. Theor. Biol.* 25, 370–379.
- Savageau, M. A. (1974). Optimal design of feedback control by inhibition. *J. Mol. Evol.* 4, 139–156.
- Savageau, M. A. (1976). *Biochemical Systems Analysis: A Study of Function and Design in Molecular Biology*. Reading, MA: Addison-Wesley.
- Savageau, M. A. (1998). Development of fractal kinetic theory for enzyme-catalysed reactions and implications for the design of biochemical pathways. *Biosystems* 47, 9–36.
- Savageau, M. A., Coelho, P. M., Fasani, R. A., Tolla, D. A., and Salvador, A. (2009). Phenotypes and tolerances in the design space of biochemical systems. *Proc. Natl. Acad. Sci. U.S.A.* 106, 6435–6440. doi: 10.1073/pnas.0809869106
- Savageau, M. A., and Voit, E. O. (1987). Recasting nonlinear differential equations as s-systems: a canonical nonlinear form. *Math. Biosci.* 87, 83–115.
- Voit, E. O. (ed.) (1991). *Canonical Nonlinear Modeling. S-System Approach to Understanding Complexity*, New York, NY: Van Nostrand Reinhold.
- Voit, E. O. (1992). Symmetries of s-systems. *Math. Biosci.* 109, 19–37.
- Voit, E. O. (2000). *Computational Analysis of Biochemical Systems: A Practical Guide for Biochemists and Molecular Biologists*. Cambridge, UK: Cambridge University Press.
- Voit, E. O., and Savageau, M. A. (1986). Equivalence between s-systems and volterra systems. *Math. Biosci.* 78, 47–55. doi: 10.1016/0025-5564(86)90030-1

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2016 Löwe, Kremling and Marin-Sanguino. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



Modeling the Metabolism of *Arabidopsis thaliana*: Application of Network Decomposition and Network Reduction in the Context of Petri Nets

Ina Koch^{1††}, Joachim Nöthen^{1†} and Enrico Schleiff^{2*}

¹ Department of Molecular Bioinformatics, Institute of Computer Science, Cluster of Excellence "Macromolecular Complexes", Goethe-University Frankfurt, Frankfurt am Main, Germany, ² Department of Biosciences, Institute of Molecular Biosciences, Molecular Cell Biology of Plants, Cluster of Excellence "Macromolecular Complexes", Goethe-University Frankfurt, Frankfurt am Main, Germany

OPEN ACCESS

Edited by:

Julio Vera González,
University Hospital Erlangen, Germany

Reviewed by:

Vincenzo Manca,
University of Verona, Italy
Guido Santos Rosales,
University of La Laguna, Spain

*Correspondence:

Ina Koch
ina.koch@bioinformatik.uni-frankfurt.de
Enrico Schleiff
schleiff@bio.uni-frankfurt.de

[†] Shared first authorship.

Specialty section:

This article was submitted to
Systems Biology,
a section of the journal
Frontiers in Genetics

Received: 18 April 2016

Accepted: 06 June 2017

Published: 30 June 2017

Citation:

Koch I, Nöthen J and Schleiff E (2017)
Modeling the Metabolism of
Arabidopsis thaliana: Application of
Network Decomposition and Network
Reduction in the Context of Petri Nets.
Front. Genet. 8:85.
doi: 10.3389/fgene.2017.00085

Motivation: *Arabidopsis thaliana* is a well-established model system for the analysis of the basic physiological and metabolic pathways of plants. Nevertheless, the system is not yet fully understood, although many mechanisms are described, and information for many processes exists. However, the combination and interpretation of the large amount of biological data remain a big challenge, not only because data sets for metabolic paths are still incomplete. Moreover, they are often inconsistent, because they are coming from different experiments of various scales, regarding, for example, accuracy and/or significance. Here, theoretical modeling is powerful to formulate hypotheses for pathways and the dynamics of the metabolism, even if the biological data are incomplete. To develop reliable mathematical models they have to be proven for consistency. This is still a challenging task because many verification techniques fail already for middle-sized models. Consequently, new methods, like decomposition methods or reduction approaches, are developed to circumvent this problem.

Methods: We present a new semi-quantitative mathematical model of the metabolism of *Arabidopsis thaliana*. We used the Petri net formalism to express the complex reaction system in a mathematically unique manner. To verify the model for correctness and consistency we applied concepts of network decomposition and network reduction such as transition invariants, common transition pairs, and invariant transition pairs.

Results: We formulated the core metabolism of *Arabidopsis thaliana* based on recent knowledge from literature, including the Calvin cycle, glycolysis and citric acid cycle, glyoxylate cycle, urea cycle, sucrose synthesis, and the starch metabolism. By applying network decomposition and reduction techniques at steady-state conditions, we suggest a straightforward mathematical modeling process. We demonstrate that potential steady-state pathways exist, which provide the fixed carbon to nearly all parts of the network, especially to the citric acid cycle. There is a

close cooperation of important metabolic pathways, e.g., the *de novo* synthesis of uridine-5-monophosphate, the γ -aminobutyric acid shunt, and the urea cycle. The presented approach extends the established methods for a feasible interpretation of biological network models, in particular of large and complex models.

Keywords: systems biology, Petri net, *Arabidopsis thaliana* metabolism, model verification, network reduction, transition invariant, common transition pairs, invariant transition pairs

1. INTRODUCTION

Arabidopsis thaliana (*A. thaliana*) is a popular model organism in plant biology (Van Norman and Benfey, 2009). *A. thaliana* was the first plant with sequenced genome (Arabidopsis-Genome-Initiative, 2000), and a large mutant collection (Sessions et al., 2002; Alonso et al., 2003) provides the optimal base for genetic and physiological analysis of this model system. It further is characterized by a short generation time, small plant size, diploid genetics, and a large number of offspring, which is of high advantage for breeding for research (Meinke et al., 1998; Koornneef and Meinke, 2010 and references therein). Most of the current information on plant metabolism is based on this model system (Lunn, 2007). Beside the academic interest in understanding the metabolism of plants, there is a broad interest to improve nutritional quality of crops and agricultural productivity to generate phytopharmaceutical substances and to increase the production of nutraceutical biomolecules and pharmaceutical proteins of commercial interest (Hur et al., 2013). The development of network models based on these investigations (Dersch and Beckers, 2016) represents a useful approach for the analysis and simulation of the metabolism. Nowadays, the Path2Models database contains 125 models for *A. thaliana* (Büchel et al., 2013). To develop a mathematical model, the single reactions are typically extracted from databases such as AraCyc (Mueller et al., 2003; Zhang et al., 2005), and/or KEGG (Kanehisa et al., 2008).

Various models of the metabolism of *A. thaliana* exist. Several approaches concern gene regulatory networks (Lucas and Brady, 2013). For example, a method based on Bayesian networks was formulated to investigate root cell differentiation (Bruex et al., 2012), and a network, focusing on stress response in leaves, was developed (Hickman et al., 2013). The power of these networks has been documented, for example, by identification of new factors involved. Using a recently developed statistical linear regression technique, novel genes were detected which are involved in the mucilage biosynthesis (Vasilevski et al., 2012).

In this paper, we focus on recent research results of metabolic modeling approaches of *A. thaliana*. The first steady-state *Metabolic Flux Analysis* (MFA) maps for *A. thaliana* were introduced in 2008 (Williams et al., 2008), using a heterotrophic cell suspension grown under two different oxygen concentrations as experimental data source. The results suggest a possible alteration of metabolite abundance without changes in the balance between respiratory and biosynthetic flux or a major rearrangement of the network. Based on this study, further investigations have been followed, in particular on the flux in the pentose phosphate pathway (Masakapalli et al., 2010). Three

new models were derived, which differ in the compartmental organization of the pentose phosphate pathway. The measured data fit to each of the three models in an acceptable manner, which necessitate further investigations in addition to the MFA. This underlines the problems of metabolic flux analysis (Masakapalli et al., 2010). Several genome-scale flux models have been developed for *A. thaliana*, for example, by Poolman et al. (2009), de Oliveira Dal'Molin et al. (2010), Radrich et al. (2010), and Mintz-Oron et al. (2012).

The *Poolman model* is based on the AraCyc database (Mueller et al., 2003) and was automatically extracted. It consists of 1,253 metabolites and 1,406 reactions, involving the production of biomass components, such as nucleotides, amino acids, lipid, starch, and cellulose in the proportion experimentally observed in a heterotrophic suspension culture. After the removal of reactions that are not necessary to maintain a steady state, 855 reactions remain. Additionally, the authors provide a steady-state model of 232 reactions that exhibit only nonzero flux values.

Similarly, de Oliveira Dal'Molin et al. (2010) automatically extract a core reaction system called AraGEM from the KEGG database to generate a model. The model is compartmentalized into cytosol, mitochondrion, plastid, peroxisome, and vacuole. Additional information is manually integrated from databases, such as AraPeroX (Reumann et al., 2004) and SUBA (Arabidopsis Subcellular Database, Heazlewood et al., 2007). The model contains 1,567 reactions and 1,748 metabolites. A two-dimensional annotation provides links from the reactions to the genes. The authors modify 36 reactions by manual curation to give a consistent stoichiometry. To achieve a desired functionality, they introduce 148 biomass drains and inter-organelle transporters.

The *Radrich model* represents a high-quality core consensus model obtained by a systematic comparison of compounds and reactions between the databases KEGG and AraCyc. Various levels of consensus lead to three submodels of different quality, the core model of 753 reactions and 914 metabolites with highest reliability, the intermediate model of 1,388 reactions and 1,248 metabolites, and the complete model of 2,315 reactions and 2,328 metabolites. The core model is the intersection between the two databases, containing every metabolite and reaction, which is present in both databases. In the intermediate model, every reaction is present, for which either all educts or all products are part of the core model. The complete model is the union of both databases. The *Mintz-Oron model* was semi-automatically derived from KEGG and AraCyc, including compartmental information from the Arabidopsis Subcellular Database, SUBA (Heazlewood et al., 2007), and tissue-specific localization data from the literature. It consists of 1,363 reactions and 1,078

metabolites. The authors predicted a total of 942 out of 1,363 inspected reactions to take place in every tissue. These reactions include reactions of the primary metabolite pathways, such as the glycolysis, the pentose phosphate pathway, and the fatty acid, nucleotide, and amino acid metabolism. The model is validated by comparison of experimental data with predicted flux values. All presented models use the databases AraCyc and/or KEGG as starting points for automatic network generation.

We chose Petri nets (PNs) as mathematical formalism for network construction, analysis, and simulation. PNs are mainly applied in computer science, for example, to model distributed systems. Many sound, rigorous analysis and simulation methods of PNs have been evolved over decades for various applications (Billington and Reisig, 1996). Some of these concepts and algorithms have been successfully applied to biochemical systems, including metabolic networks (Koch et al., 2005), signal transduction networks (Sackmann et al., 2006), gene regulatory networks (Matsuno et al., 2000), protein complex assembly (Bortfeldt et al., 2011), and combinations of them (Grunwald et al., 2008; Koch et al., 2011).

To motivate the work, we aimed to develop a consistent metabolic model that reflects the steady-state condition and the basic dynamic behavior. Thus, the model becomes suitable for rigorous mathematical analyses and can serve as basis for quantitative analyses. Our work was motivated by a hand-built PN model for the metabolism of barley *Hordeum vulgare*, integrating biochemical, physiological, proteomic, and genomic data derived from the literature and databases (Grafahrend-Belau et al., 2009). We followed the paper's strategy and built a model for *A. thaliana* by successively adding metabolites and reactions from the literature (Nöthen, 2009). We systematically expanded the PN to develop a new model based on the current experimental knowledge on the metabolism of *A. thaliana* and to investigate the structural and dynamic properties of the model. To check a metabolic model for consistency and correctness, we considered special network properties (Heiner and Koch, 2004). For biochemical systems, a correct biological interpretation of submodules at steady state was mainly applied. The computation of such submodules was based on minimal semi-positive transition invariants (TIs), also known as elementary modes (Schuster and Hilgetag, 1994). For the definition, see Section 2.1. Minimal semi-positive integer solutions were of interest (Schrijver, 1998; Koch and Ackermann, 2013) leading to a complexity that does not allow the computation of all solutions, even if the power of supercomputers is used. Today, several alternative methods for the exploration of the inherent flux states of large systems exist (Koch and Ackermann, 2013).

The paper is organized as follows. In the Section 2, we introduce Petri nets and describe the network verification and the network reduction techniques we used, including the common transition pairs and invariant transition pairs. In Section 3, we give the complete PN and the reduced PN model of the central metabolism in *A. thaliana*. We explain an example for a network reduction and for a TI, and discuss the *Maximal Common Transition sets*, covering the analyses of functional modules. The supplementary file Table 1.pdf contains the metabolites, the reactions, and the output reactions, including the literature

references. The supplementary file Table 2.pdf contains the tables, indicating each reduction step. The supplementary file Supplementary Material Data Sheet 1 includes all Petri net models.

2. MATERIALS AND METHODS

In the following, we briefly describe the methods for network reduction and network decomposition given in the Petri net formalism.

2.1. Petri Nets

Petri nets (PNs) are based on a concept of communication of automation originally introduced by Carl Adam Petri (Petri, 1962) in his dissertation in 1962 for mathematical modeling of causal systems with concurrent processes. PN provide a flexible, well-defined, mathematical formalism for various modeling types, ranging from qualitative to quantitative modeling. Here, we introduce the basic definitions and notations that are necessary to understand the paper. For a more detailed introduction, see Murata (1989), Baumgarten (1996), and Koch et al. (2011).

The first biological application of PNs was published in 1993 by Reddy et al. (1993). PNs has been used to model various biochemical systems, such as metabolic networks (Koch et al., 2005), signal transduction networks (Sackmann et al., 2006), gene regulatory networks (Matsuno et al., 2000), protein complex assembly (Bortfeldt et al., 2011), and combinations of them (Grunwald et al., 2008). The PN formalism has been also used for stochastic modeling, applying the Gillespie's algorithm (Gillespie, 1977), and for kinetic modeling, applying mass action kinetics and/or Michaelis-Menten kinetics. For a review, see Koch et al. (2011).

Petri nets are directed, bipartite, labeled graphs. They consist of two disjunctive sets of vertices, P and T , respectively. The elements in P are the *places* graphically represented as circles, and the elements in T are the *transitions* graphically represented as rectangles. Places stand for the passive system's elements, for example, chemical compounds, metabolites, proteins, and protein complexes. Instances of them are represented by *tokens*, which define discrete entities. Since the movement of tokens realizes the dynamics of a PN, we introduce the concept of tokens in Section 2.1.1. Transitions are the active system's elements, for example, chemical reactions, degradation processes, and complex assembly processes. Places and transitions are connected by directed, labeled edges. Edges between vertices of the same type are not allowed. Usually, the edges are labeled or weighted, respectively, by integer numbers. In the following, we call a directed, bipartite graph *net* or *net graph*, see Definition 2.1. To define a Petri net, we extend that definition, compare Definition 2.2.

Definition 2.1. Net or net graph.

A net or net graph is a triple $N = (P, T, F)$ with

$P \cap T = \emptyset$ and

$F \subseteq (P \times T) \cup (T \times P)$.

We call the elements of P and T places and transitions,

respectively. They are the vertices or nodes of the net. We call the elements of the flow relation, F , edges or arcs.

Regarding a net vertex, we define two sets of neighbor vertices, the set of all pre-vertices $\bullet x := \{y \mid (y, x) \in F\}$ and the set of all post-vertices $x \bullet := \{y \mid (x, y) \in F\}$. Accordingly, we write for a set of *pre-places*, $\bullet t$, for a set of *post-places*, $t \bullet$, for a set of *pre-transitions*, $\bullet p$, and for a set of *post-transitions*, $p \bullet$.

We call edges that are going in two opposite directions as *read arcs* or *test arcs*. Using read arcs, we can model, for example, catalytic reactions, where the catalyst is necessary to activate the transition (reaction), but will not be consumed, when the transition takes place. A PN without read arcs is a *pure PN*.

Definition 2.2. Petri net. A Petri net or Place/Transition net or P/T net is a six-tuple $Y = (P, T, F, K, W, M_0)$, if

- (P, T, F) is a net,
- $K : P \rightarrow \mathbb{N} \cup \{\infty\}$ (capacities of the places, possibly infinite),
- $W : F \rightarrow \mathbb{N}$ (edge weights or label of the edges), and
- $M_0 : P \rightarrow \mathbb{N}_0$ (initial marking) with $\forall p \in P : M_0(p) \leq K(p)$.

2.1.1. The Dynamics of P/T Nets

The dynamics of a net is performed by movable objects called *tokens* which are located on the places and will be removed according to the *firing rule*. If N is a P/T net, a mapping of the $M : P \rightarrow \mathbb{N}_0$ with $\forall p \in P : M(p) \leq K(p)$ is called a marking of N . We graphically represent the number of tokens of a place, $M(p)$, under a marking, M , by $M(p)$ dots (tokens) on the place, p . We write the capacities $\neq \infty$ and edge weights $\neq 1$ at the places and edges, respectively. The token distribution over all places define a certain state of the net. $\mathcal{M}(N)$ is the set of all markings of N . In the following, let M be a fixed marking of N . The number of the tokens can be restricted by the capacity of the place. In most biological applications, the capacity is set to infinite. Additionally, we introduce *logical* vertices to get an improved layout. A logical vertex has copies of the same name in the graphical representation of the model. Using logical vertices, we can draw PNs in a clearly-arranged way. For example, in metabolic networks, ATP participates in many reactions. This leads to many crossing edges if we model only one place for ATP. To avoid these crossing edges, we copy the place for ATP and mark it as logic place.

In this paper, we consider the classical *Place/Transition PN* (P/T net). The firing rules do not include time relations. A transition *fires*, or for biochemical networks, a reaction takes place, if the transition is *activated* or *has concession*, i.e., if the pre-places carry at least as many tokens as indicated by the weights of the corresponding edges and if the capacity of the post-places is large enough as indicated by the corresponding edge weights. At the moment of firing, the tokens of the pre-places will be consumed, and the tokens on the post-places will be produced, both according to the corresponding edge weights. A new system's state is achieved. The numbers of tokens and the edge weights implement quantitative properties although the system is still discrete. If always at least one transition of the PN is activated, the PN contains no deadlock and is called to be *deadlock-free*.

Definition 2.3. Activated transition. A transition, $t \in T$, is activated or has concession under M , written as $M \xrightarrow{t}$, if

- $\forall p \in \bullet t : M(p) \geq W(p, t)$,
- $\forall p \in t \bullet : M(p) \leq K(p) - W(t, p)$.

We say that t fires from M to M' and write $M \xrightarrow{t} M'$, if t is activated under M , and M' arises from M by removal of tokens from the pre-places and production of tokens on the post-places according to the corresponding edge weights:

$$M'(p) = \begin{cases} M(p) - W(p, t), & \text{if } p \in \bullet t \setminus t \bullet, \\ M(p) + W(t, p), & \text{if } p \in t \bullet \setminus \bullet t, \\ M(p) - W(p, t) + W(t, p), & \text{if } p \in t \bullet \cap \bullet t, \\ M(p) & \text{otherwise.} \end{cases} \quad (1)$$

Transitions without pre-places are always activated. We call them *input transitions*. Accordingly, we call transitions without post-places *output transitions*. Input and output transitions were used to model the interface to the system's environment. **Figure 1** illustrates an example of a pure PN and its firing behavior.

2.1.2. Linear Invariants

Using linear invariants, we can define specific dynamic properties of the network. These properties are valid in every system state. They can be used for model verification, but also for model decomposition and network reduction. For PNs, we define the *place invariants* (PI) for the passive part and the *transition invariants* (TI) for the active part. The definitions of both invariant types are based on the incidence matrix, we also know as stoichiometry matrix for metabolic networks. This $m \times n$ matrix, C , for m places and n transitions indicates for each place, $p_i \in P$, the change in the number of tokens, Δm , when a transition, $t_j \in T$, fires.

Definition 2.4. Incidence matrix. Let N be a P/T net. The corresponding incidence matrix C is defined by: $\forall 1 \leq i \leq m, 1 \leq j \leq n$

$$C_{ij} = \begin{cases} W(t_j, p_i), & \text{if } (t_j, p_i) \in F \setminus F^{-1}, \\ -W(p_i, t_j), & \text{if } (p_i, t_j) \in F \setminus F^{-1}, \\ W(t_j, p_i), & \text{if } (t_j, p_i) \in F \cap F^{-1}, \\ 0 & \text{otherwise.} \end{cases} \quad (2)$$

Note that for a PN without read arcs, we can remove $\setminus F^{-1}$ and the third condition. Contrary, for all cases, the third case, $C_{ij} = W(t_j, p_i)$, if $(t_j, p_i) \in F$, would be sufficient, if we consider W as complete mapping to $(P \times T) \cup (T \times P)$ and set the weights of non-existing edges to 0.

We developed the software tool *MonaLisa* especially designed for PN applications to biological systems (Einloft et al., 2013). Additionally to an intuitive editor, the tool provides many useful analysis functions based on PNs as well as on graph theory. It allows for classical discrete modeling as well as for stochastic modeling (Balazski et al., 2015).

Definition 2.5. Place invariant (PI). Let N be a P/T net and C the corresponding incidence matrix. A place invariant of N is an m -tuple $x \in \mathbb{Z}^m$ with $C^T x = 0$.

Definition 2.6. Transition invariant (TI). Let N be a P/T net and C the corresponding incidence matrix. A transition invariant of N is an n -tuple $y \in \mathbb{Z}^n$ with $Cy = 0$.

We are interested in the minimal, semi-positive, integer solutions; *integer*, because we are working at discrete level, *semi-positive* because there is no interpretation of negative solutions, and *minimal* because such equation systems can have an infinite number of solutions. In the following, we consider minimal, semi-positive, integer PIs and TIs, writing shortly PIs and TIs, respectively. For discussion of the computational problem, see Schrijver (1998) and Koch and Ackermann (2013).

The support of a vector, x , is represented by the non-zero entries of the vector written as $\text{supp}(x)$. An invariant x is called *minimal*, if its support does not contain the support of any other invariant z , i.e.,

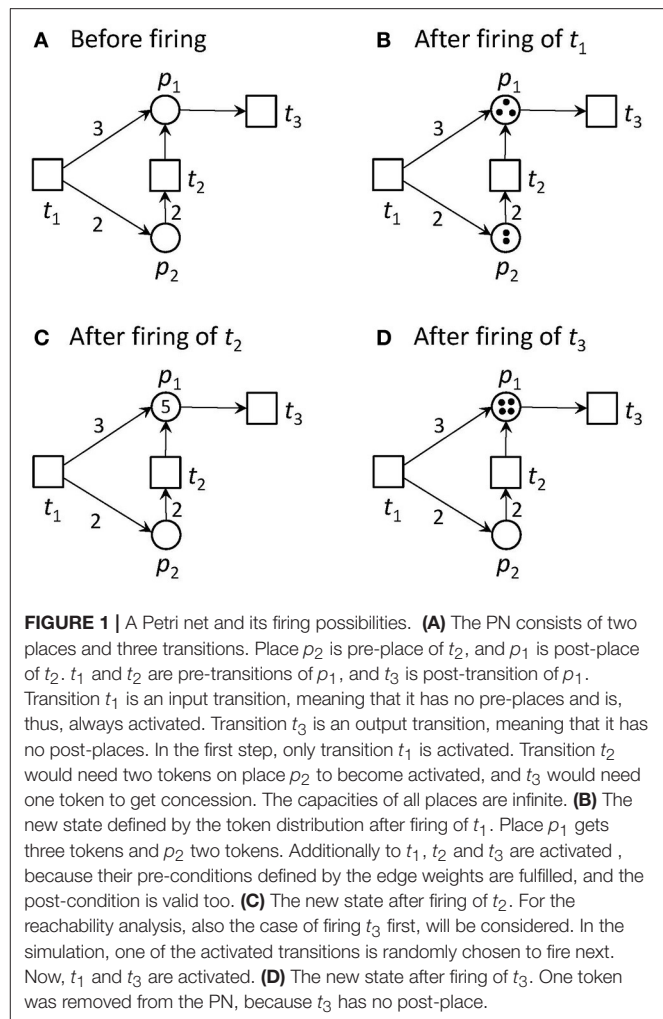
$$\nexists \text{ invariant } z : \text{supp}(z) \subset \text{supp}(x), \quad (3)$$

and the greatest common divisor of all non-zero entries of x is one. Whereas PIs reflect a token or substance conservation, the TIs describe basic functional modules of the system's dynamics at steady state (Lautenbach, 1973; Schuster and Hilgetag, 1994; Schuster et al., 2002). These functional modules have to be checked for their biological correctness. The firing of the transitions of a TI in the given frequency reproduces the initial state. Thus, a TI is also called a *Parikh* vector. If every place or every transition, respectively, is member of at least one PI or TI, respectively, the PN is *covered by PIs (CPI)* or *covered by TIs (CTI)*, respectively. The CTI property indicates the network's completeness or consistency. A transition which is not member of at least one TI does not contribute to the systems behavior. Thus, it could be removed without influencing the overall systems behavior. A TI or PI, respectively, defines a connected subnet, consisting of its support, the support's pre- and post-places or pre- and post-transitions, respectively, and all edges in between.

TIs can be classified according to the type of the participating transitions. The classes of TIs are motivated by the involved input and output transitions. We consider the following types of TIs:

- trivial: a reversible reaction, which is split into a forward and a backward transition.
- INOUT: contains at least one input transition and one output transition.
- IN: contains an input transition, but no output transition.
- OUT: contains an output transition, but no input transition.
- CYC: contains neither an input nor an output transitions forming a cycle in the model.

The explanations for some of the different types of invariants are intuitive. In a metabolic PN, containing reversible reactions, it is not possible to avoid trivial TIs. TIs of the type INOUT represent pathways through the network, a succession of consecutive biochemical reactions, transforming given educts (metabolites produced by input transitions) to the corresponding products



(metabolites consumed by output transitions). TIs of the types IN or OUT contain only input or output transitions, respectively. These TIs emerge from internal production or consumption, respectively, of secondary metabolites. TIs of type CYC contain neither input nor output transitions at all. Their firings form cycles in the PN.

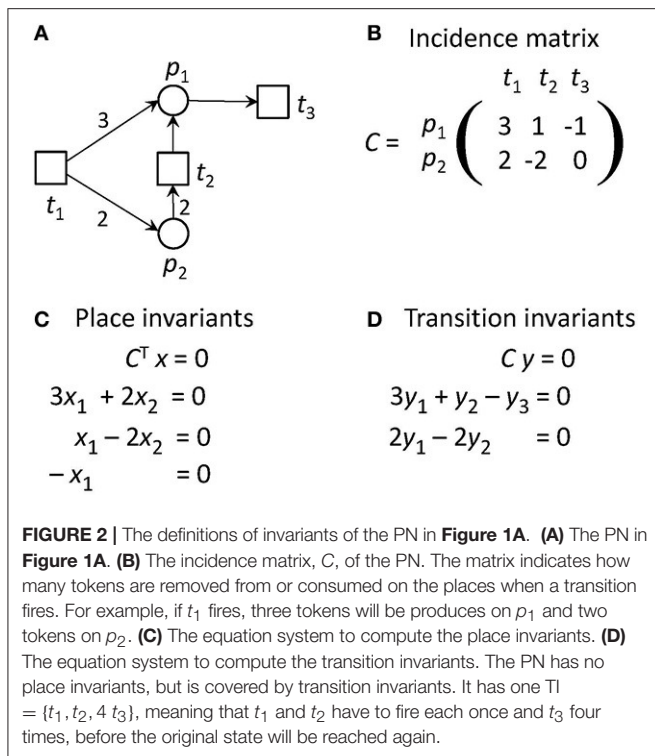
Figure 2 gives the incidence matrix and the equation systems for PI and TI computation of the PN example of Figure 1.

2.1.3. Maximal Common Transition Sets (MCT-Sets)

To support the biological interpretation of TIs, we group the transitions into *Maximal Common Transition sets (MCT-sets, MCTS)* by their occurrence in the minimal TIs: $\forall i, j \in \{1, \dots, m\}$ the transitions, t_i and t_j , are grouped into the same MCT-set, if and only if they participate in exactly the same minimal TIs, i.e., all TIs x hold:

$$\chi_{\{0\}}(x_i) = \chi_{\{0\}}(x_j), \quad (4)$$

whereas $\chi_{\{0\}}$ denotes the characteristic function, binary indicating whether an argument is equal to zero. This grouping



leads to maximal sets of transitions, whereat each set of transitions ϑ holds:

$$\forall x \in X: \quad \vartheta \subseteq \text{supp}(x) \vee \vartheta \cap \text{supp}(x) = \emptyset, \quad (5)$$

whereas X denotes the set of all minimal TIs, x .

This grouping represents an equivalence relation in T , the set of transitions, which leads to a partition of T . The equivalence classes ϑ correspond to the MCT-sets. MCT-sets define also subnets as TIs, but they have not to be connected. The subnets defined by MCT-sets are disjunctive. They represent a further decomposition method of large biochemical networks into rather small subnets at steady state, which often correspond to functional units. Each of the MCT-sets may represent a building block with a special biological meaning.

2.2. Network Verification

Network verification is a crucial part in the process of model construction. For biological PNs, beside structural and behavioral properties such as the connectivity (Heiner and Koch, 2004; Koch et al., 2011), another important property is the biological interpretability of the TIs (Heiner and Koch, 2004; Koch et al., 2005; Grunwald et al., 2008) and the MCT-sets (Sackmann et al., 2006; Grafahrend-Belau et al., 2008).

The CTI property results from the TIs. This property can be seen as one indicator for the consistency and completeness of a network. It is of particular importance that a biological network is CTI, because this property ensures that each reaction may contribute to the basic system behavior (Koch and Heiner, 2008) while the steady state of the system is preserved.

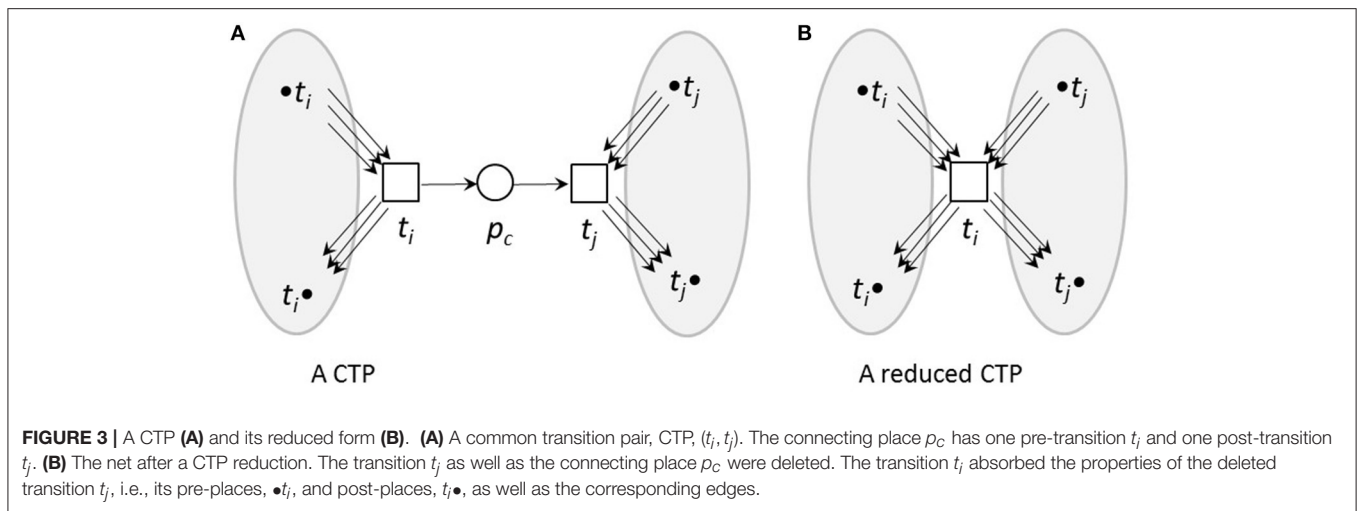
We assume that metabolic networks reach a steady state. This steady-state assumption is reflected in the computation of the TIs. A TI represents a set of reactions whose enzymes ensure the steady-state condition. Only one missing reaction would lead to a disturbance of the steady state of the system. Moreover, each TI should represent a functional module in the overall network dynamics. The verification or interpretation for a biological meaning of each functional module, i.e., of each TI, represents a method to verify large networks with regard to their correctness. In this context, the computability of all TIs is a big problem, for even middle-sized networks, i.e., consisting of some hundreds of vertices and thousands of edges (Ackermann and Koch, 2011). To handle this problem, we applied network reduction techniques.

2.2.1. Network Reduction

Network reduction methods should reduce the complexity of the system, conserving main properties and the main behavior of the network. Reduction techniques are essential for network verification and analysis of big, complex systems. In computer science, the question for the correctness of algorithms, their running times, and other theoretical aspects of the reduction process are of great interest (Arnborg et al., 1993). Reduction has been employed in the analysis of complex networks in the Petri net community during the last decades. Various methods for the structural reduction of series of transitions and places have been developed (Lee-kwang et al., 1987). In this reduction process, surrounding vertices were summarized under conservation of special PN properties. Other techniques rely on the sharing properties of two or more vertices. Parallel transitions or places are connected to the same sets of pre-places and post-places or pre-transitions and post-transitions, respectively. Then, these parallel structures can be merged (Murata, 1989). Other methods try to integrate deadlock-avoiding policies, for example, in flexible manufacturing systems (Uzam, 2004). Beside these structural reduction methods, there exist also dynamical techniques for PNs (Berthelot, 1987).

In systems biology, reduction techniques are used at the topological level as well as at steady-state level. All these techniques help to verify and analyze biochemical networks. Investigations on metabolic networks (Reddy et al., 1993) apply reduction techniques of parallel structures. One approach establishes a matrix-based method to study the hierarchical organization of metabolic networks (Ravasz et al., 2002). To reduce the size of the matrix, non-branching pathways in the metabolic network of *Escherichia coli* are replaced by single reactions, resulting in a decrease of complexity.

Here, we used topological reduction techniques as well as steady-state reduction techniques. We considered topological reduction techniques that preserve the CTI-property (Ackermann et al., 2012). To reduce chains of transitions, we defined *Common Transition Pairs (CTPs)* as two transitions, t_i and t_j , connected by a place called a *connecting place*, p_c , see **Figure 3**. We deleted the transition t_j as well as the connecting place, p_c . The transition, t_i , absorbed the properties of the deleted transition, t_j , i.e., all its pre-places, $\bullet t_i$, and post-places, $t_i \bullet$, as well as the corresponding connecting edges.



To reduce reversible reactions, we defined *Invariant Transition Pairs (ITPs)* as two transitions, t_i and t_j , representing the forward and backward reaction, respectively, of a reversible reaction and connecting two places, p_i and p_j , see **Figures 4, 5** for an example. We deleted the ITP (t_i, t_j) , as well as the place p_j . The place p_i absorbed the properties of the deleted place p_j , i.e., all its pre-transitions, $\bullet p_i$, and post-transitions, $p_j \bullet$, as well as the corresponding connecting edges and the markings.

To indicate a reduction, we changed in each reduction the names of the reactions and the metabolites. If two reactions were merged in a CTP reduction, we provided a new name for the merged transition, e.g., “ $ctp(t_i + t_j)$ ”. If we identified, for example, the two reactions *E151* and *E152* connected by metabolite 87 as a potentially reducible CTP, we connected all edges from reaction *E152* to reaction *E151*, removed metabolite 87 and reaction *E152*, and renamed reaction *E151* to $ctp(E151 + E152)$.

If the reduction was based on an ITP, we renamed the merged metabolites. According to the CTP reduction, we connected the names of the two metabolites in the same way preceded by an ITP. If we reduce, for example, the reactions *E24_f* and *E24_b*, which connect the metabolites 76 and 50, by an ITP reduction, we would connect all edges from metabolite 50 to metabolite 76, remove the reactions *E24_f* and *E24_b* and the metabolite 50, and rename metabolite 76 to $itp(76 + 50)$.

Additionally, we reduced parallel reactions, i.e., if two transitions had the same input place and the same output place, and the same weights on the corresponding edges, we combined both transitions into one without changing the net behavior (Murata, 1989). We adopted this rule for the analysis of biochemical networks (Reddy et al., 1993). We used this technique to further reduce the complexity of the model by reducing parallel pairs of transitions, but we applied it to those with only one pre-place and one post-place and edges of weight 1. The reduction procedure was recursive. It was possible that, e.g., the place $itp(76 + 50)$ could be identified as an ITP with another place in the next step. Each step of the reduction could be followed by the nomenclature and produced a timescale of

reductions, compare with the Tables 1–9 in the supplementary file Table 2.pdf.

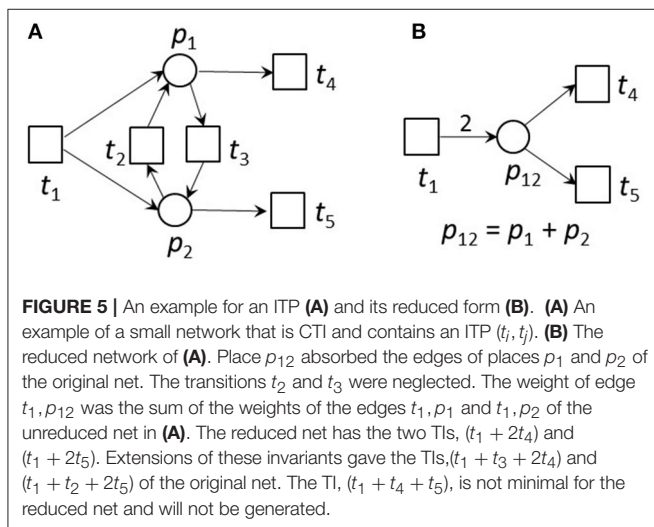
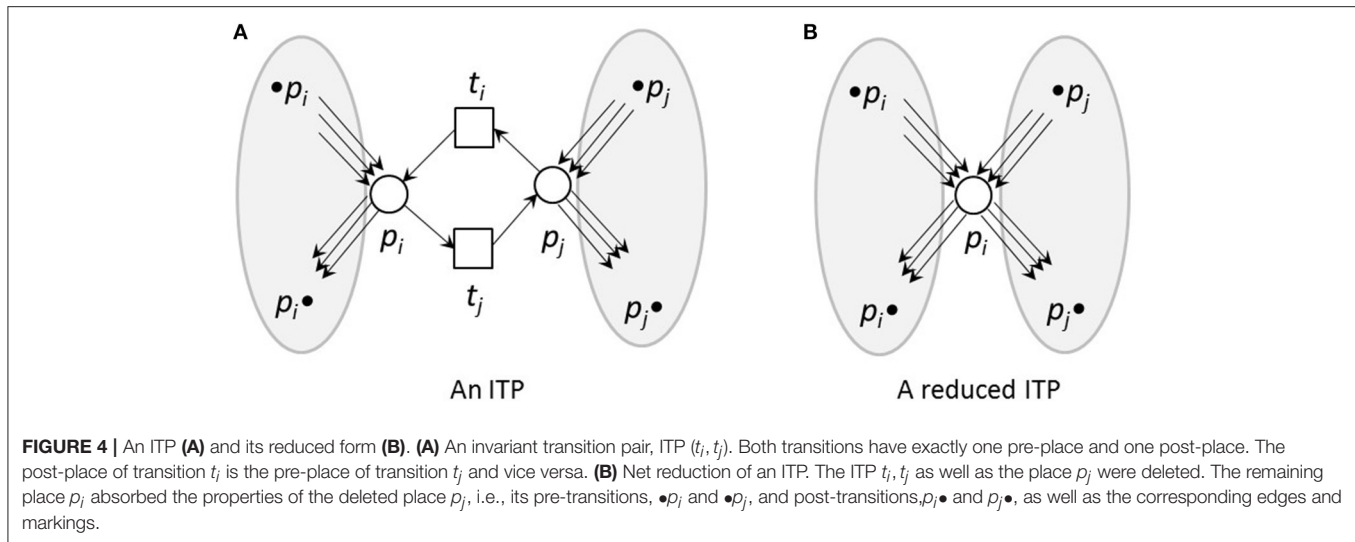
A place or transition that had no connections left was removed from the model. It was possible to remove the just reduced CTP, i.e., the transition $t_i \in T$, if the transition pair, (t_i, t_j) , was additionally a reversible reaction for all other places connected to it. When this situation appeared, each ingoing edge for each connected place would cancel an outgoing edge of the same place, and, therefore, cancel all edges from and to $t_i \in T$. After an ITP reduction, it was possible that another ITP was removed, if it connects the same p_i and p_j . For every other potential ITP, the reduction of the first ITP resulted into two transitions connected to the reduced place $p_i \in P$ with an ingoing and outgoing edge. We then removed these transitions, because they were no longer connected to the model.

3. RESULTS AND DISCUSSION

3.1. The Petri Net Model

We developed the PN model based on the literature, i.e., each reaction (transition) is experimentally proven. The complete PN model consists of 134 metabolites and 243 reactions, which are connected via 572 edges. **Figure 6** on the top gives a schematic illustration of the model. The metabolites within the PN were numbered. Their names are listed in the Supplementary file Table 1.pdf. The reactions, including the literature references, are compiled in the Tables 2–13 in the Supplementary file Table 1.pdf. The interface to the environment was modeled by 29 reactions indicated by an *IN* or *OUT* in the prefix of their names. The input reactions create the substrates glycine (metabolite = place 18), D-fructose (place 63), D-galactose (place 73), coenzyme A (CoA, place 93), acetyl-CoA (place 92), ammonia (place 29), or citrulline (place 94). In turn, the output reactions create 22 metabolites (see Table 14 in the Supplementary file Table 1.pdf). 18 output transitions were connected to sink metabolites.

Four metabolites were connected to both, an input and output transition, forming external metabolites (CoA, ammonia, acetyl-CoA, and citrulline). The biosynthesis of CoA was not modeled,



and thus, the precursor of CoA, pantothenate (Raman and Rathinasabapathi, 2004), is not present in the PN. Similarly, acetyl-CoA is a product of the β -oxidation of fatty acids in *A. thaliana* (Fulda et al., 2002) which is not part of the model as well. Nitrate and ammonia were taken up by roots and transported to photosynthetic tissues (Chalot et al., 2006). In leaves, ammonium is the primary nitrogen source for glutamine synthesis in the cytosol or chloroplasts, while nitrogen is either stored in vacuoles or converted to ammonium for further processing. For intracellular transport of ammonium between mitochondria and chloroplasts, a so-called ornithine-citrulline shuttle is proposed (Linka and Weber, 2005), which is also discussed to represent an efficient transport system for carbon dioxide from mitochondria to chloroplasts in form of citrulline. We modeled the external setting of citrulline to dissolve an occurring deadlock, consisting of the metabolites 94 (citrulline), 27 (L-arginino-succinate), 28 (arginine), and 1 (ornithine). Considering the four metabolites

that serve as input and output, the PN consumed 3 input metabolites to produce 18 output metabolites.

For this complete PN, we were not able to compute the TIs. To verify the PN we first manually divided the complete PN into four smaller subnetworks of known biological meaning, for which we could separately compute the TIs. We chose four subnets of biological relevance according to the literature: (1) the sucrose (Figure 6, blue), (2) the citrate (red), (3) the shikimate (yellow), and (4) the UTP subnet (green). The names of the subnets represent the key metabolites of each subnet. The sucrose subnet was supplied with D-fructose and D-galactose as input from the environment, while the other five input metabolites were fed into the citrate subnet. In turn, nine metabolites were exported from the sucrose subnet, five from the citrate subnet, five from the shikimate subnet, and three from the UTP subnet.

3.1.1. The Sucrose Subnet

The sucrose subnet consists of 44 metabolites and 103 reactions (Figure 7). The subnet contains seven additional input reactions and 22 additional output reactions to generate a sufficient outlet of products. The metabolites involved in these reactions link the sucrose network with the other parts of the PN. 18 output transitions were connected to sink metabolites.

The sucrose subnet consists of four pathways, the Calvin cycle (Figure 7, blue), the sugar metabolism (red), the glycolysis (yellow), and the starch metabolism (green). The Calvin cycle is central for carbon fixation in plants, while the sugar metabolism accomplishes the synthesis and degradation of sucrose (metabolite 66) and UDP-glucose (metabolite 82). The glycolysis (yellow, orange) appeared to be disconnected in the layout, but the reactions of the glycolysis were combined via the logical place 44 (β -D-fructose 6-phosphate). However, only a part of the glycolysis pathway belongs to the sucrose network—the reaction cascade from D-glucose (metabolite 51) to glycerate 3-phosphate

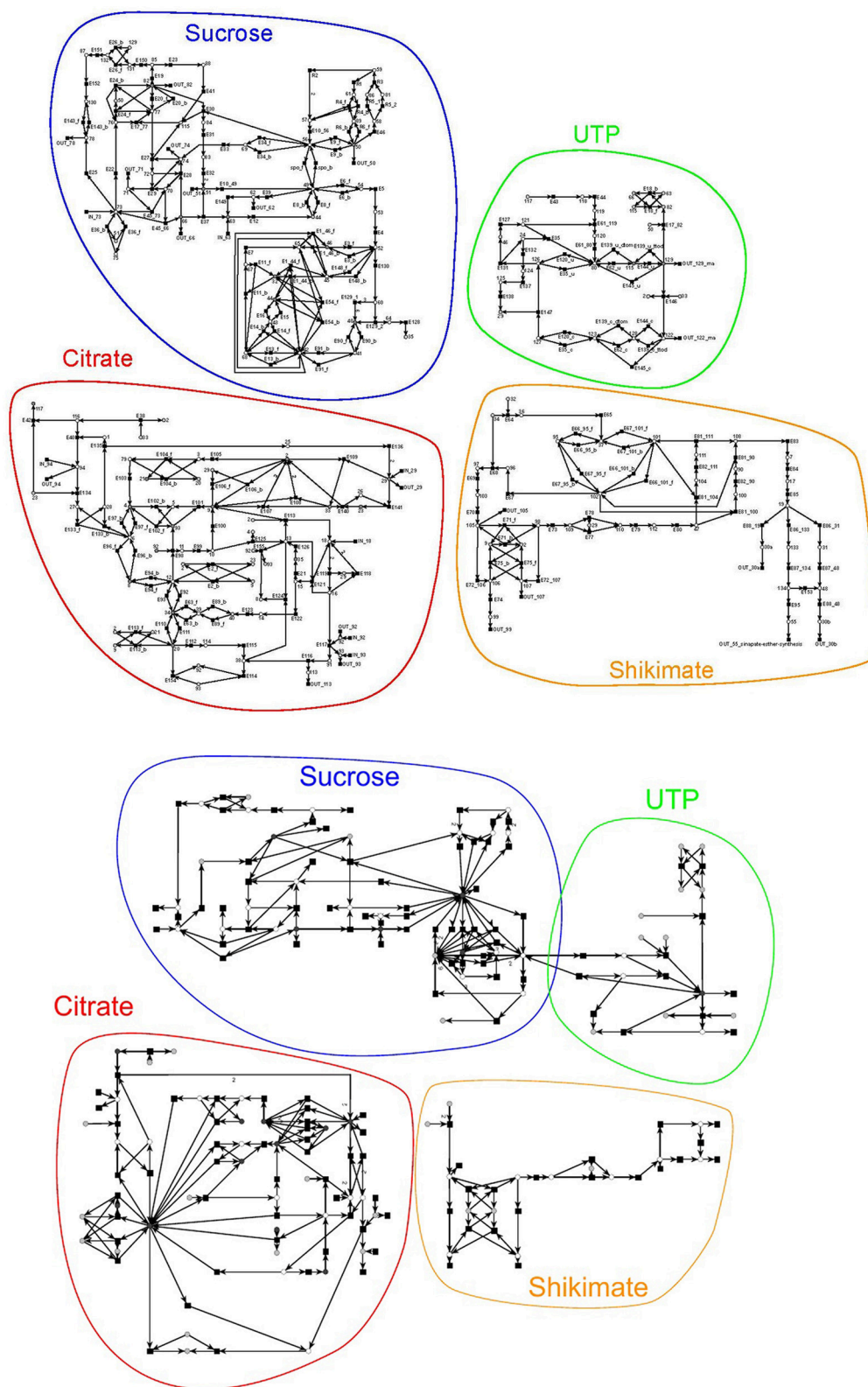
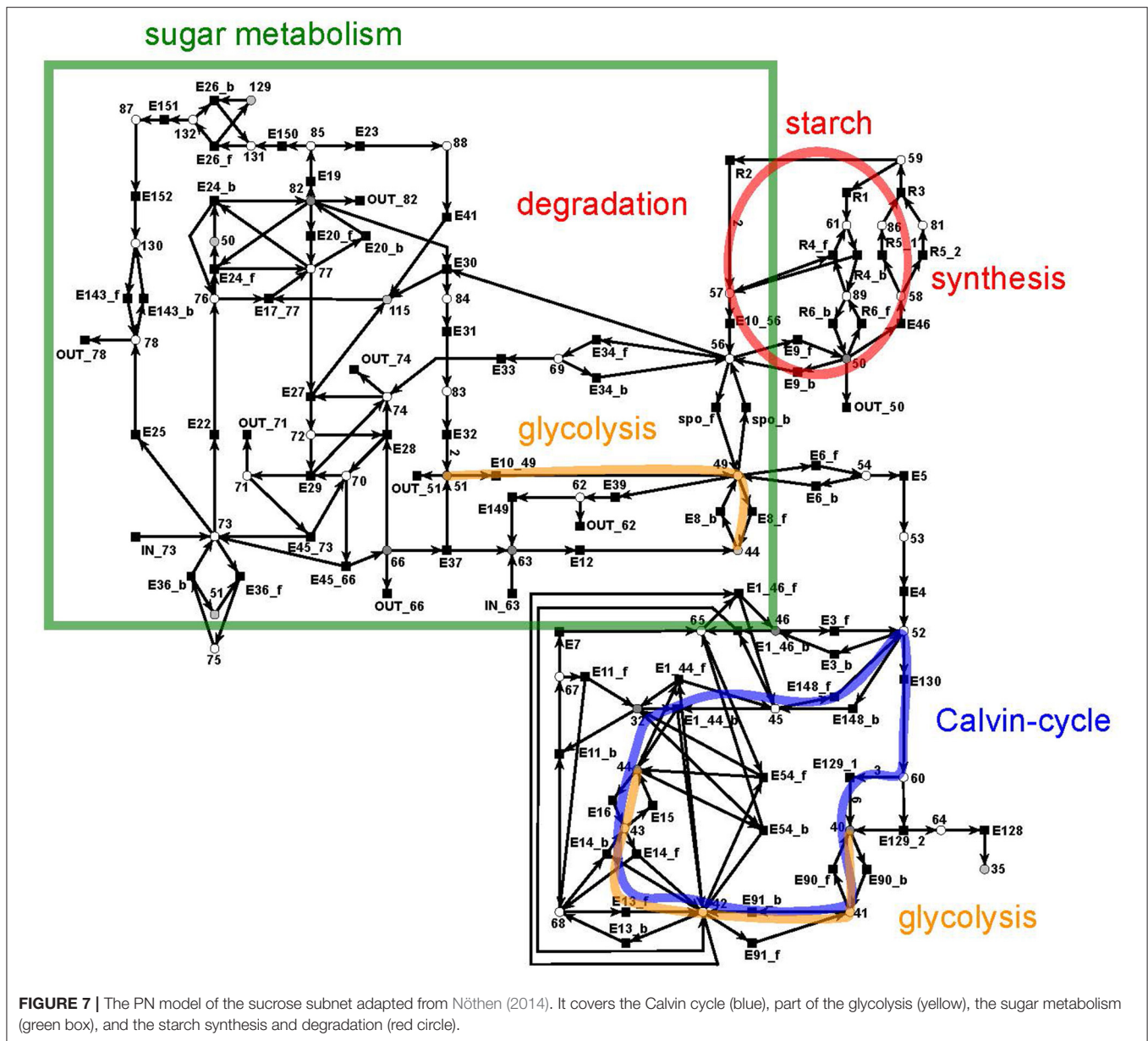


FIGURE 6 | Schemas of the original and reduced PN model adapted from Nöthen (2014). It is a coarse-grained representation of the metabolism of *A. thaliana*. The original model (on the top) consists of 134 metabolites and 242 reactions connected via 569 edges and the reduced model (on the bottom) of 60 metabolites, 131 reactions, and 329 edges. The models were manually subdivided into four subnets, the sucrose, the UTP, the citrate, and the shikimate net. The subnets are highlighted in the same color and name.

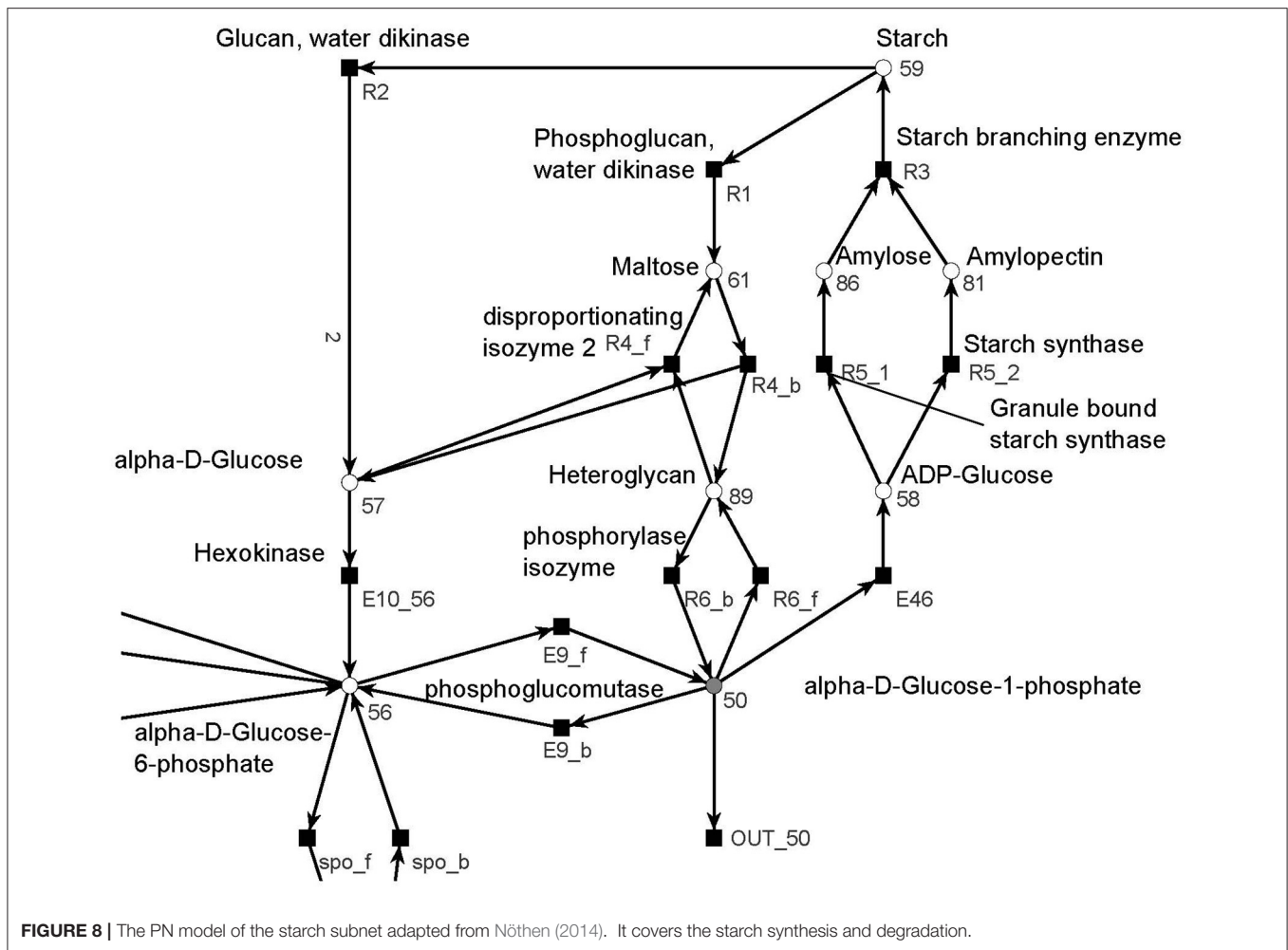


(metabolite 40). In turn, the other part of the glycolysis pathway is integrated in the citrate subnet. Thus, a transfer metabolite was required for the complementation of this cycle.

3.1.2. The Starch Metabolism

The metabolism became obvious by inspecting the synthesis and degradation of starch in detail based on **Figure 7** (red). The priming, chain-addition polymerization, polymer degradation, irreversible poly-condensation, and granule formation of starch are complex enzymatic processes. The mechanisms underlying these processes are not yet fully understood for *A. thaliana* (Szydlowski et al., 2009). While constructing the PN (**Figure 8**), we did not account for the diverse structures of starch macromolecules and for the chain length distribution. Thus,

we lumped the diversity of starch macromolecules to a single unique metabolite named starch (metabolite 59). This simplification was supported by a biologically intuitive view of the coarse-grained structure of the net. We adopted previously presented suggestions (Kossmann and Lloyd, 2000; Guy et al., 2008; Fettke et al., 2009) to model the starch metabolism as a pure, deadlock-free PN. To indicate deviations of reactions from the activity of single enzymes we replaced the prefix *E* (for enzyme) by *R* (for reaction) in the names of such reactions. More precisely, in **Figure 9**, two molecules of ADP-glucose (metabolite 58) form one molecule of starch (metabolite 59) via amylose (metabolite 86) and amylopectin (metabolite 81) (Kossmann and Lloyd, 2000). The breakdown of one unit of starch produces two units of α -D-glucose (metabolite 57) or, alternatively, via maltose (metabolite 61) only one unit of α -D-glucose and one unit of



heteroglycan (metabolite 89) (Fettke et al., 2009). Heteroglycan is freely convertible to α -D-glucose 1-phosphate (metabolite 50). Also α -D-glucose can be converted via α -D-glucose 6-phosphate (metabolite 56) to α -D-glucose 1-phosphate (metabolite 50). α -D-glucose 1-phosphate is represented by a logical place (gray-filled circle). Thus, the metabolite α -D-glucose 1-phosphate can be produced and consumed by several reactions also outside the sucrose subnet. The reaction, *E46*, transforms α -D-glucose 1-phosphate to ADP-glucose, which can be consumed for further production of starch.

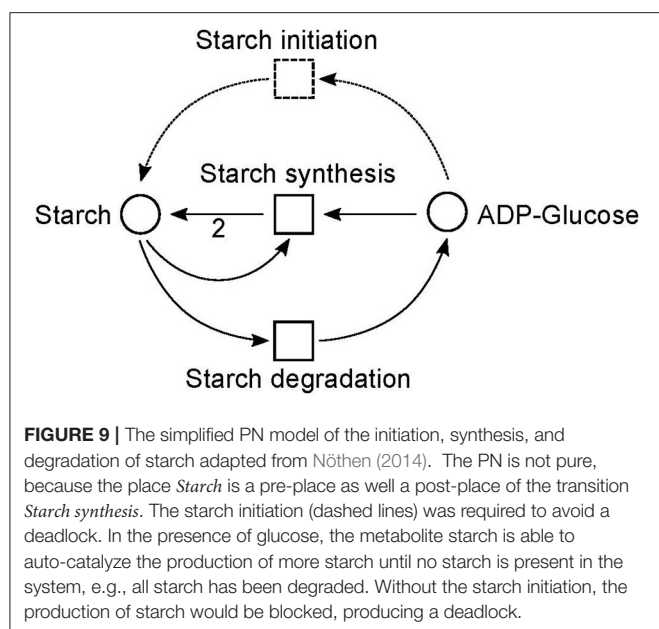
Due to the polymer character of the starch molecule, it was difficult to model its synthesis and degradation in detail. There is literature (Kossmann and Lloyd, 2000; Lu and Sharkey, 2006; Guy et al., 2008; Reiter, 2008; Fettke et al., 2009) available that describes the reaction cascade of the starch metabolism without special focus on the polymeric structure of starch. We put special effort into the curation of the thermodynamic feasibility of the starch pathway by an adaption of the stoichiometric parameters in the cascade in such a way that no substance was created or consumed (comparable to the procedure in Poolman et al., 2009). This thermodynamic feasibility was supported by the two cyclic TIs,

one for each degradation pathway, which together covered the starch metabolite.

3.1.3. The Citrate Subnet

The citrate subnet consists of 43 metabolites and 79 reactions (Figure 10). The subnet contains two additional input and five additional output reactions that connect the citrate network with the other parts of the PN.

This subnet specifies the biosynthesis of glutamate (metabolite 2) and glutamine (metabolite 33) (blue), the citric acid cycle (red), and the glyoxylate cycle (green), part of the biosynthesis of uridine 5'-phosphate (violet), and completes the pathway of glycolysis (yellow). Glutamine and glutamate play a central role in the transfer of ammonia (metabolite 29). The citric acid cycle is part of the energy metabolism in aerobic species. The glyoxylate cycle (glyoxylate: metabolite 13) plays a major role in the anabolic synthesis of carbohydrates. The reactions of glycolysis complement the part of the cycle integrated in the sucrose subnet by converting glycerate 3-phosphate (metabolite 40) to pyruvate (metabolite 20) (yellow). The biosynthesis of uridine 5'-phosphate (uridine mono phosphate, UMP) is only



partially present in this subnet and was completed in the UTP subnet (purple).

3.1.4. The Shikimate Subnet

The shikimate pathway (Figure 11) consists of 39 metabolites and 46 reactions. It indicates the synthesis of shikimate (metabolite 102: red) as precursor for the synthesis of aromatic amino acids, e.g., phenylalanine (metabolite 109: green) (Herrmann and Weaver, 1999). In turn, shikimate and phenylalanine are precursors for phenylpropanoids (blue) which, in turn, are precursors for the lignin (metabolites 30a and 30b) synthesis. The synthesis of lignin is e.g., explicitly included as a part of the biomass function in a genome scale flux balance model of maize (*Zea mays*) (Saha et al., 2011). We divide lignin into two compounds, 30a and 30b, representing guaiacyl lignin and syringyl lignin, respectively. Guaiacyl lignin is believed to be synthesized from coniferyl alcohol (metabolite 19) while syringyl lignin is produced from sinapyl alcohol (metabolite 48) (Humphreys and Chapple, 2002).

3.1.5. The UTP Subnet

The UTP subnet is depicted in (Figure 12, red). It consists of 25 metabolites and 29 reactions. The connection with the other parts of the PN was achieved by three additional output reactions (Table 14 in the Supplementary file Table 1.pdf).

The subnet contains the completion of biosynthesis of UMP (Figure 12, red) that is initiated within the citrate subnet by consuming N-carbamoyl-L-phosphate (metabolite 117) to produce UMP (metabolite 80). The second major part of this subnet describes the interconversion of the pyrimidine ribonucleotides, i.e., the connection between UMP, UDP, and UTP and between CMP, CDP, and CTP, and the synthesis of CTP from UTP and the degradation of CMP to UMP (green).

3.2. The Reduced Petri Net Model

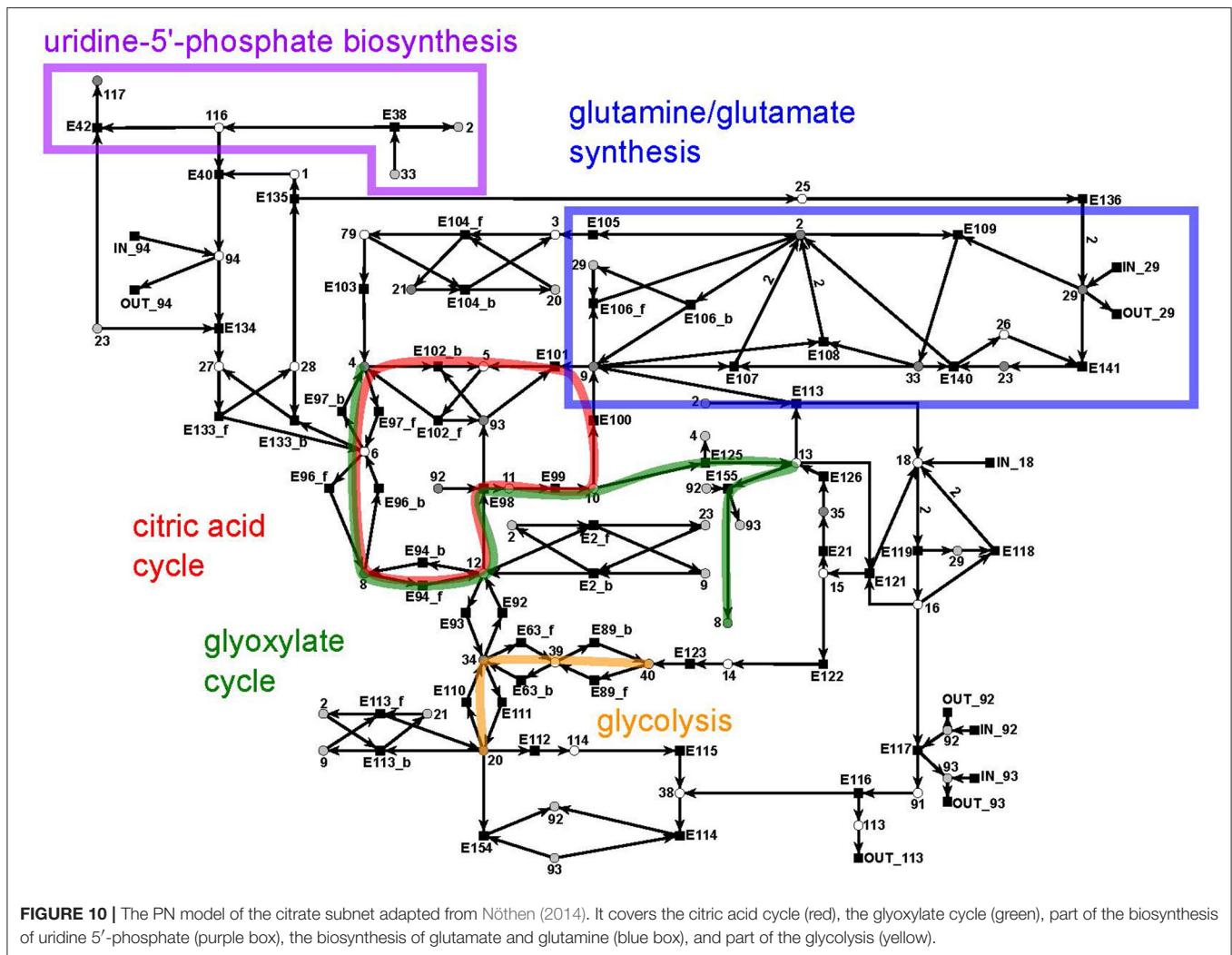
3.2.1. Removal of Metabolites

To limit the network complexity as far as possible, all small metabolites like water and carbon dioxide, catalytic substances like ions, and cofactors like NAD/NADH and the energy-providing compounds like ATP were omitted from the model. Hereby, the model was reduced by 19 metabolites and 274 edges. Studies have proven that hubs are less conserved than compounds which connect different modules (Guimera and Amaral, 2005), suggesting that a model without cofactors and small metabolites is more suitable than a model which lacks essential connections between the different biological modules. A complexity reduction by removing such smaller metabolites was preferred to a restriction of the overall network size. Several of the small metabolites, namely CoA and UDP, remained in the network due to suggestions of our coworkers (Schleiff, 2010, personal communication). To prove that the reduced network was still a real-world network, we compared the distributions of vertex degrees of the reduced network with that of the network based on the AraCyc database. After excluding all metabolites with a vertex degree greater than 74, we could show that the distribution of vertex degrees of the AraCyc network was very similar to the one of the PN model. This indicated that the reduced PN model was also a real-world model.

We applied ITPs (Invariant Transition Pairs) and CTPs (Common Transition Pairs) using an extended version of a previously published algorithm (Ackermann et al., 2012; Nöthen, 2014). We searched first for CTPs and then for ITPs in a parallel running implementation. If any of these structures were encountered, the net was reduced and followed by a search for a CTP. We considered the following cases: (1) there are CTPs that as well form ITPs. This is the case, if a place is connected to the rest of the network only via an ITP-forming transition pair. Then, the algorithm reduces the ITP first. (2) A special case of an ITP is given, if a certain metabolite is connected to an input and output transition. The PN model contains several external metabolites which fulfill these special ITP rules.

86 CTPs, 62 ITPs, and 2 parallel reduction steps were performed. All steps were listed in the supplementary file Table 2.pdf in the Tables 1–9. The number of metabolites was reduced from 134 to 60, the number of reactions from 243 to 131, and the number of edges from 572 to 329.

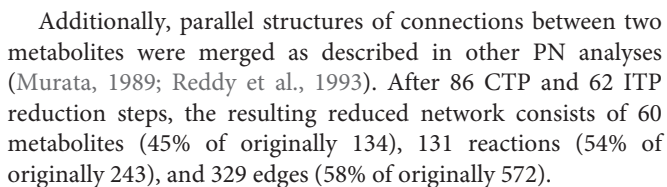
Figure 6 depicts in a schematic way the original PN model (on the top) and the model after reduction (on the bottom). We exemplarily demonstrated the interpretation of the results of the reduction procedure, choosing a place following from a series of ITP reductions. The place represented a conglomeration of substances involved in the Calvin cycle, the glycolysis, and the citric acid cycle. It combined the compounds succinate (metabolite 4), fumarate (metabolite 6), malate (metabolite 8), oxalacetate (metabolite 12), pyruvate (metabolite 20), phosphoenolpyruvate (metabolite 34), glycerate 2-phosphate (metabolite 39), glycerate 3-phosphate (metabolite 40), glycerate 1,3-bisphosphate (metabolite 41), glyceraldehyde 3-phosphate (metabolite 42), and dihydroxyacetone phosphate (metabolite 68) into the same place. These reduction steps additionally removed the transitions E13, E14, E15, E90, and E91,



while the transitions *E11* and *E14* remained in the reduced PN. The first four metabolites were part of the citric acid cycle. Out of the remaining seven metabolites which were part of the glycolysis, there were five, namely glyceralate 2-phosphate (metabolite 39), glyceralate 3-phosphate (metabolite 40), glyceralate 1,3-bisphosphate (metabolite 41), glyceraldehyde 3-phosphate (metabolite 42), and dihydroxyacetone phosphate (metabolite 68), which were also part of the Calvin cycle.

In plastids of *A. thaliana*, the glycolysis and the Calvin cycle shared several enzymes (Peltier et al., 2006): phosphoglycerate kinase (transition *E90*, reversible), glyceraldehyde 3-phosphatedehydrogenase (transition *E91*, reversible), triosephosphate isomerase (transition *E13*, reversible), seduheptulosebisphosphate aldolase (transitions *E11* and *E14*, both reversible), and fructose bisphosphatase (transition *E15*, reversible by transition *E16*). The sharing of these reactions suggested that the participating compounds were shared as well. In this case, the shared metabolites would be: glyceralate 3-phosphate (metabolite 40), glyceralate 1,3-bisphosphate (metabolite 41), glyceraldehyde 3-phosphate (metabolite 42),

D-fructose 1,6-bisphosphate (metabolite 43), D-fructose 6-phosphate (metabolite 44), and dihydroxyacetone phosphate (metabolite 68) besides D-fructose 1,6-bisphosphate (metabolite 43) and D-fructose 6-phosphate (metabolite 44). The absence of D-fructose 1,6-bisphosphate (metabolite 43) and D-fructose 6-phosphate (metabolite 44) in the reduced place was caused by restrictions in the reduction process. A merging of places by an ITP reduction was only allowed if all involved edges had a weight of 1. This was not the case for the reactions connecting metabolite 43 with 42 and 68 (transition *E14*). The merging of parts of the citric acid cycle, namely succinate (metabolite 4), fumarate (metabolite 6), malate (metabolite 8), and oxalacetate (metabolite 12), was inspired by the reversible reactions between the compounds, which was the main idea behind an ITP reduction. The combination of this part of the citric acid cycle and the last steps of the glycolysis, producing phosphoenolpyruvate and pyruvate, was induced by the synthesis pathway of oxalacetate from phosphoenolpyruvate (Dey and Harborne, 1997). Considering these biological aspects, the reduction process made sense.

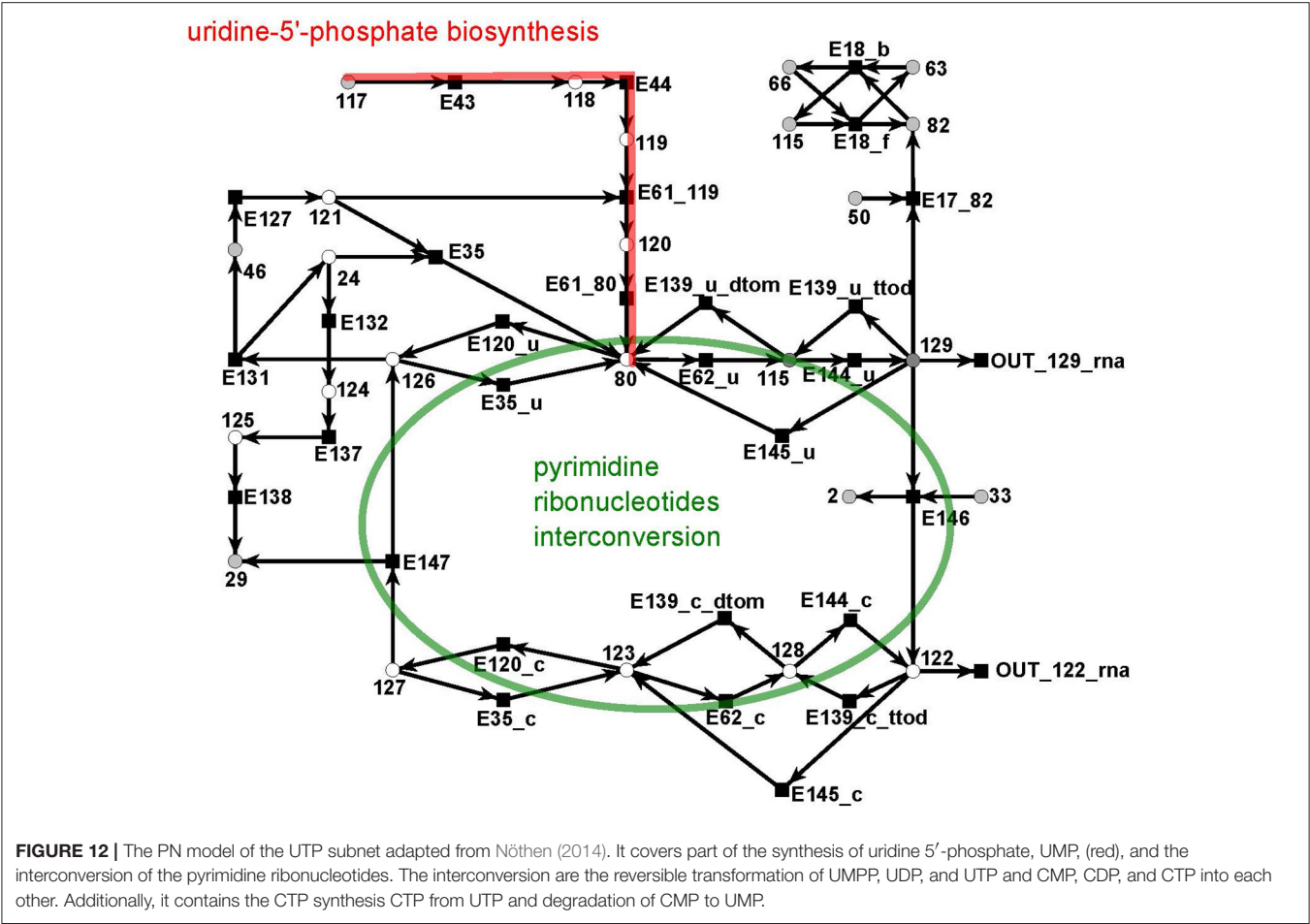


The problem of enumerating all minimal TIs can not be solved in polynomial time. The overall complexity of this task is still not clear, and the decision problem, whether two transitions occur in the same TI, is NP-complete (Acuña et al., 2010).

of TIs for the two modules, the reduced network, and for all TI types.

We could show that a mapping of reduced invariants to invariants of the unreduced net will be possible.

June 2017 | Volume 8 | Article 85



Due to deleted carbon dioxide, we modeled the Calvin cycle without an explicit input transition for carbon dioxide. All OUT TIs contained one of the two transitions modeled rubisco (E129_1 and E129_2) and thereby an implicit IN TI of carbon dioxide. This implicit modeling reduced all OUT TIs to INOUT pathways. Transition E129_1 modeled the carboxylation reaction (Calvin cycle) of rubisco, and transition E129_2 the oxygenation reaction (photorespiration) (Eckardt, 2005). All other modules did not possess OUT TIs. The OUT TIs of the reduced model were comparable to those of the sucrose subnet.

3.3.1. An Example for a Cyclic Transition Invariant

All cyclic TIs were as well artifacts, resulting from missing metabolites and cofactors. **Figure 13** illustrates an example of a cyclic TI. The modeled starch metabolism requires ATP and α -D-glucose 1-phosphate (metabolite 50) to form ADP-Glucose (metabolite 58) and pyrophosphate by transition E46 (Kossmann and Lloyd, 2000; Streb and Zeeman, 2012). ATP and pyrophosphate were not modeled, otherwise this reaction would had required ATP and produced pyrophosphate. This led to two options, (1) ATP and pyrophosphate were directly provided and removed, and (2) ATP and pyrophosphate were produced and consumed throughout the network. Both cases resulted directly

TABLE 1 | The numbers of places, transitions, and TIs in the modules and in the reduced network.

| PN size | Sucrose module | Citrate module | Reduced PN |
|----------------|----------------|----------------|------------|
| Places | 69 | 72 | 60 |
| Transitions | 132 | 125 | 131 |
| TI type | | | |
| All | 4,602 | 3,214 | 27,646 |
| Trivial | 31 | 25 | 22 |
| IN + OUT | 4,473 | 3,140 | 26,095 |
| IN | 18 | 41 | 1,298 |
| OUT | 63 | 0 | 132 |
| Cyclic | 17 | 8 | 99 |

All networks were CTI. We grouped the TIs according to the type of interface to the environment. INOUT TIs contained input and output transitions, IN TIs only input transitions, and OUT TIs only output transitions. Cyclic TIs contained neither IN nor OUT TIs. Trivial invariants were reversible reactions which were split into a forward and a backward transition.

or indirectly in the involvement of input and output transitions leading to an INOUT TI.

IN TIs were as well caused by substances, which were not modeled. In the sucrose subnet, all IN TIs were artifacts of

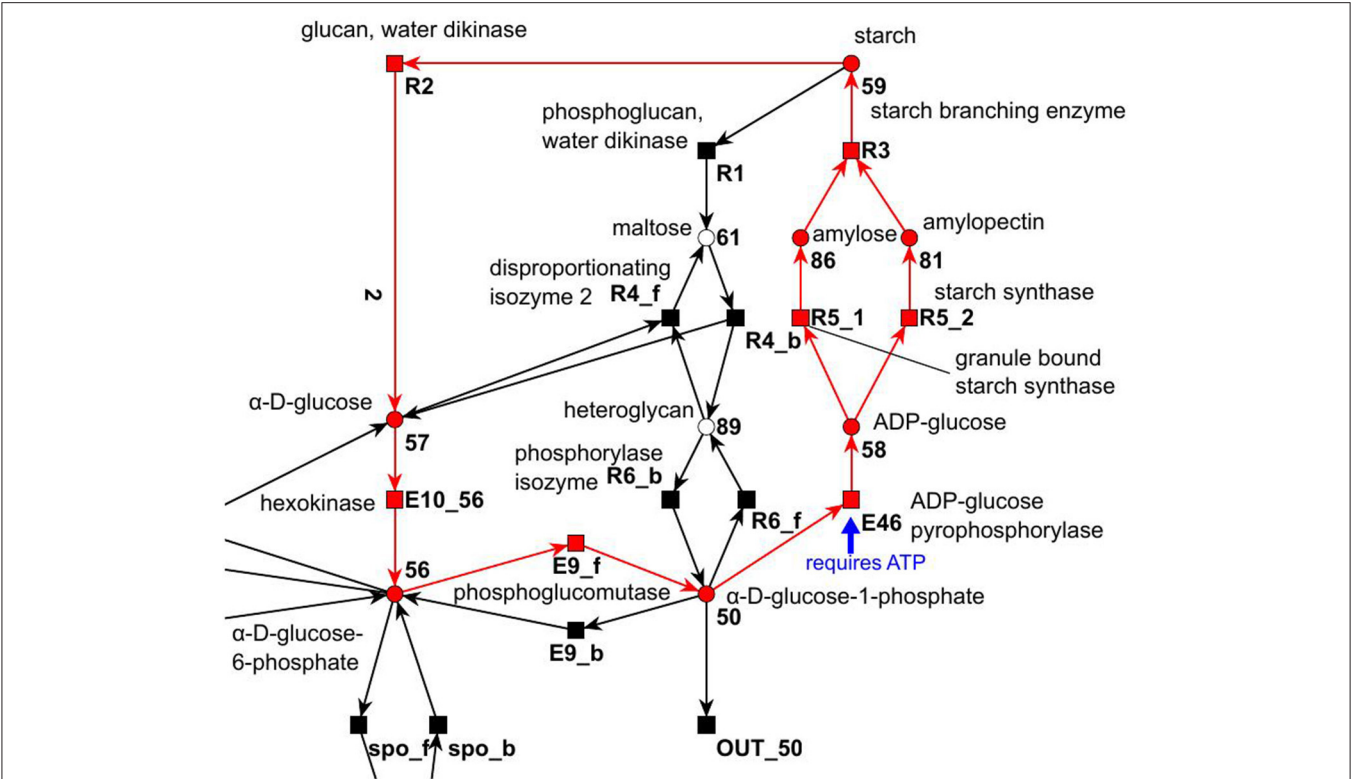


FIGURE 13 | Example of a cyclic TI that describes part of the starch synthesis and degradation. The cycle evolves out of a silent input transition. The underlying reaction of transition *E46* produces ADP-glucose and pyrophosphate from ATP and α -D-glucose 1-phosphate (Streb and Zeeman, 2012).

the modeling of the pentose phosphate pathway. Transition *E4* produces a carbon dioxide (Kruger and von Schaewen, 2003) in the pentose phosphate pathway from metabolite 53 (6-phospho gluconate). As carbon dioxide was not modeled, this had the function of a hidden export. This behavior corresponded to the OUT TIs, which were artifacts of the missing modeling of carbon dioxide as well.

3.3.2. Exemplary Inspection of a Single Transition Invariant

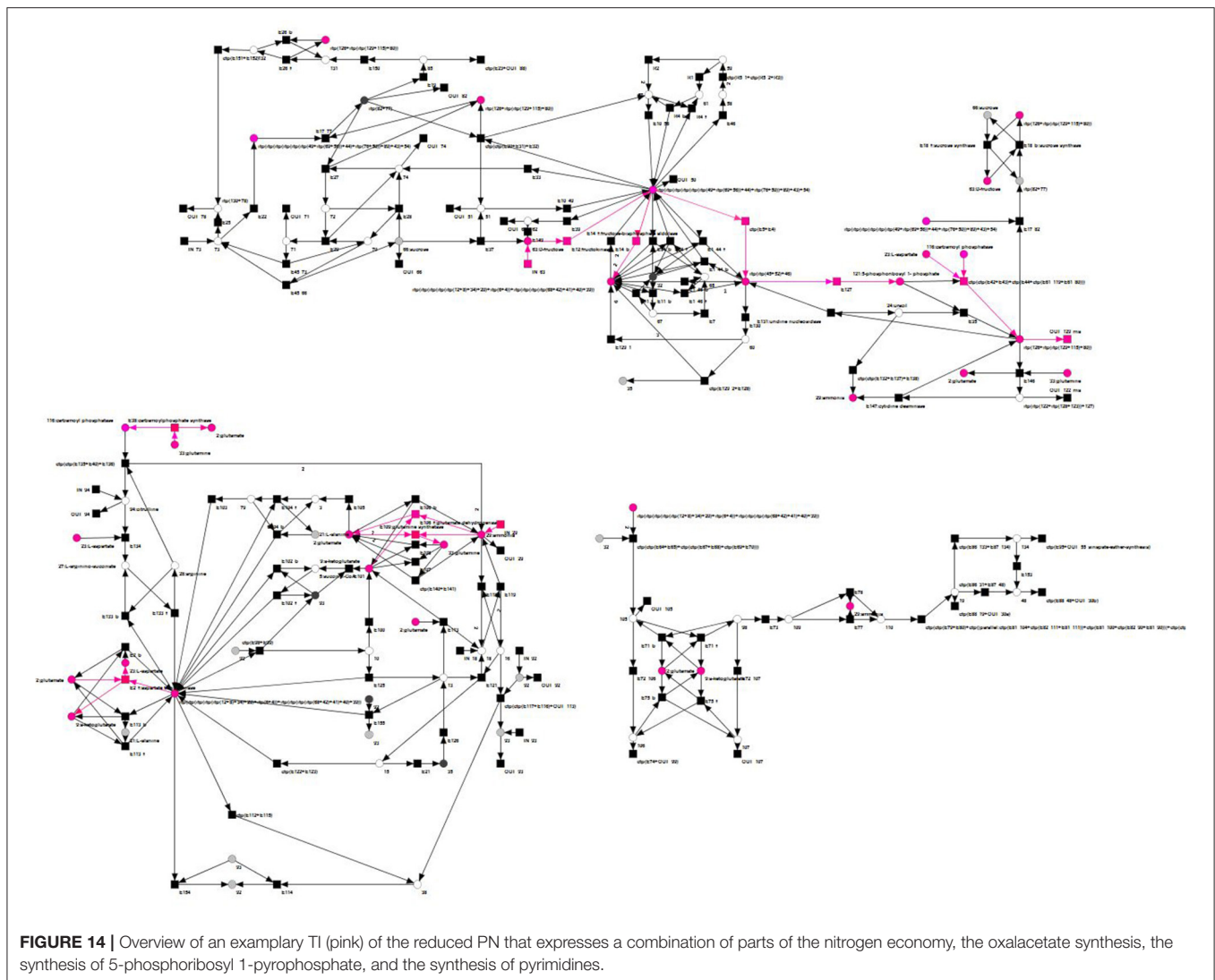
Here, we want to illustrate an exemplary inspection of a single TI of the reduced model, consisting of twelve reactions. As the reduction process merges reactions, the number of traversed reactions in the original PN could be higher. Table 2 illustrated the considered exemplary TI and listed the syntheses of substrates a reaction was involved in. Several biological pathways could be combined to form a single TI (Koch et al., 2005). The considered TI produced UTP for the synthesis of RNA from D-fructose (metabolite 63) and ammonia (metabolite 29). To biologically verify this pathway combination, we had to demonstrate that each part of the TI could be biologically explained. The product of the TI was two UTPs, which was removed from the network by the transition *OUT_129_rna*. The used compounds for this product were three D-fructoses provided by *IN_63* and four ammonia provided by *IN_29*. In plants, one UMP was synthesized from four different compounds: one bicarbonate, one glutamine,

TABLE 2 | Exemplary TI in the reduced model.

| Reaction and its firing frequency | Involved in the synthesis of |
|---|---|
| 2* <i>E2_f</i> | L-aspartate |
| <i>E14_f</i> | L-aspartate |
| 3* <i>IN_63</i> | L-aspartate, 2*5-phosphoribosyl 1-pyrophosphate |
| 3* <i>E12</i> | L-aspartate, 2*5-phosphoribosyl 1-pyrophosphate |
| 2* <i>ctp(E5 + E4)</i> | 5-phosphoribosyl 1-pyrophosphate |
| 2* <i>E127</i> | 5-phosphoribosyl 1-pyrophosphate |
| 4* <i>IN_29</i> | 2* glutamate, 2* glutamine |
| 2* <i>E106_f</i> | Glutamate |
| 2* <i>E109</i> | Glutamine |
| 2* <i>E38</i> | Uridine 5-phosphate (UMP) |
| 2* <i>ctp(ctp(E42 + E43)+</i> <i>ctp(E44 + ctp(E61_119 +</i> <i>E61_80)))</i> | Uridine 5-phosphate (UMP) |
| 2* <i>OUT_129_rna</i> | RNA |

Each reaction was listed, including the number of its occurrences in the TI. Additional information was given about the syntheses the reaction was involved in, either the synthesis of one of the substrates L-aspartate, 5-phosphoribosyl 1-pyrophosphate, glutamine, and UMP, or the final synthesis of RNA. If different subpathways were assigned to the same reaction, the respective multiplicity was mentioned (see *IN63* and *E12*).

one 5-phosphoribosyl 1-pyrophosphate, and one L-aspartate (Zrenner et al., 2006). UTP was then synthesized from UMP by adding phosphor groups. The metabolite bicarbonate was



not modeled in the PN. It was demonstrated in the following that each of the three remaining modeled compounds was synthesized and used in a biologically explainable way. Two L-aspartate were synthesized by firing of transition *E2_f* (aspartate aminotransferase) two times. The substrates of this reaction were oxalacetate and glutamate and the products L-aspartate and α -ketoglutarate (Wilkie and Warren, 1998; Graindorge et al., 2010). In the reduction process, several pathways, leading to oxalacetate, were affected, resulting in a merged metabolite, which represents various compounds. The affected pathways were the glycerate 3-phosphate synthesis, the oxalacetate synthesis, and the glycolysis.

5-phosphoribosyl 1-pyrophosphate can be synthesized from D-fructose, via D-fructose 6-phosphate and D-ribose 5-phosphate (Dey and Harborne, 1997; Buchanan et al., 2000; Berg et al., 2002; Zrenner et al., 2006). This pathway structure was represented in the TI by the reactions *E12* (fructokinase, synthesis of D-fructose 6-phosphate), *ctp(E5+E4)* (6-phosphogluconolactonase and phosphogluconate

dehydrogenase, synthesis of D-ribose 5-phosphate), and *E127* (5-phosphoribosyl 1-pyrophosphatesynthase, synthesis of 5-phosphoribosyl 1-pyrophosphate). Please note that a number of reactions in the overall synthesis of 5-phosphoribosyl 1-pyrophosphate was affected by the reduction process, and the precursor of transition *E127* combined several compounds. The remaining reactions, *E106_f* and *E109*, were required to regenerate two glutamate and two glutamine, thereby using four ammonia. Glutamine is an important part of the nitrogen metabolism of conifers (Cánovas et al., 2007), and α -ketoglutarate is a known nitrogen transporter in plants (Temple et al., 1998). As glutamate is convertible in both, glutamine and α -ketoglutarate (Aubert et al., 2001; Forde and Lea, 2007), it seems to share these nitrogen-transportation duties. In the TI, the regeneration of glutamate and glutamine, respectively, from α -ketoglutarate and glutamate and in the process, consuming ammonia, resembled this biological interpretation, see Figures 14, 15. These findings proved this TI to be a combination of parts of the nitrogen economy, the oxalacetate

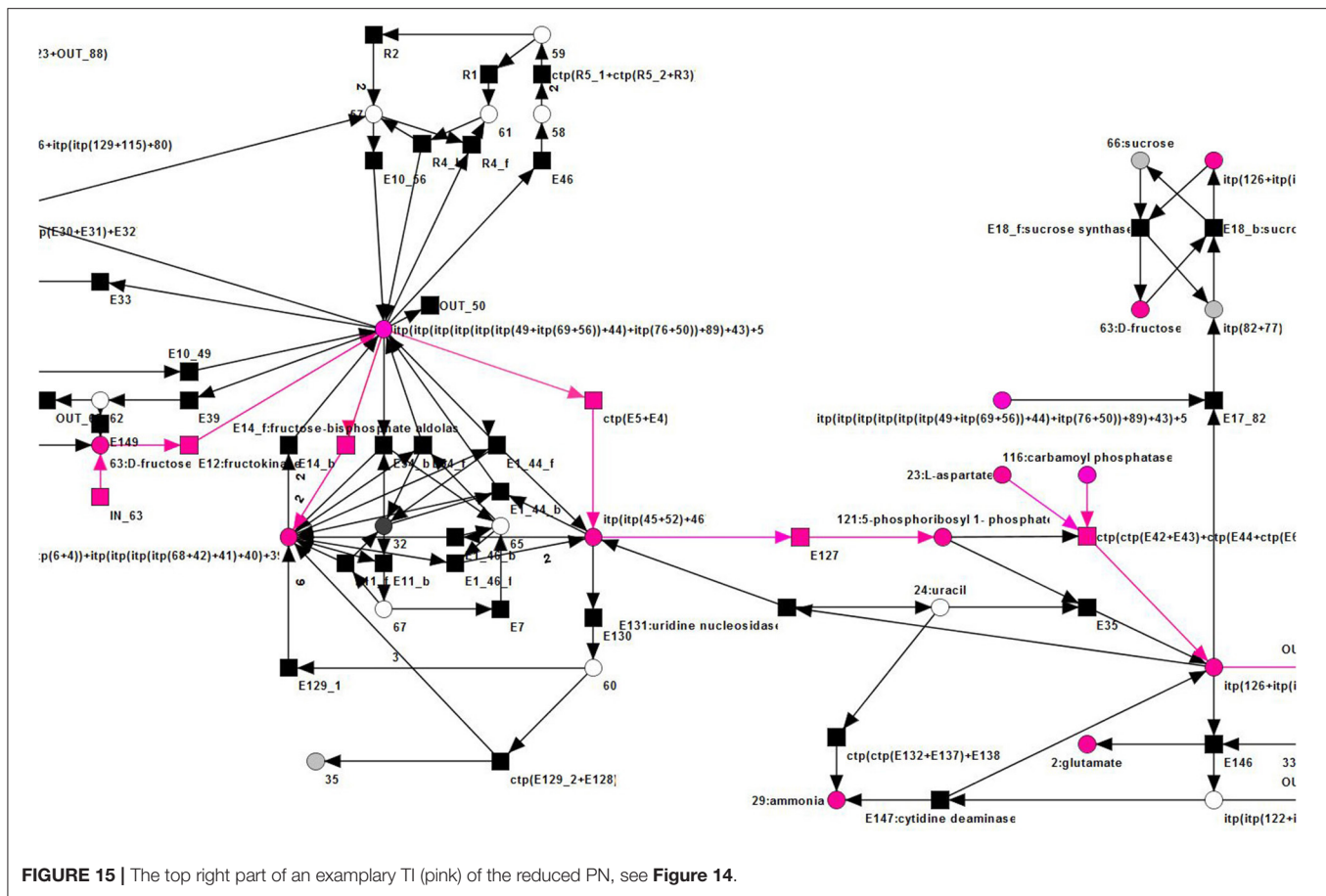


FIGURE 15 | The top right part of an exemplary TI (pink) of the reduced PN, see Figure 14.

synthesis, the synthesis of 5-phosphoribosyl 1-pyrophosphate, and the synthesis of pyrimidines. Altogether, these were the important steps to synthesize UTP (Zrenner et al., 2006). Each part of the inspected TI could be biologically interpreted, which proved the possibility of the network to model these syntheses in a steady-state sustaining manner.

3.4. MCT-Sets

MCT-sets represent the smallest biologically meaningful entities in which a network can be decomposed (Sackmann et al., 2006). We give several examples of MCT-sets of the reduced PN model and their biological counterparts. This additionally provides a possibility of comparison between the biologically motivated subnets and the reduced network.

3.4.1. MCTS 1 and the Sucrose Module

This MCTS consists of the transitions *Import117fromB/fromCit*, *E43*, *E44*, *E61_119*, and *E61_80*. All these transitions formed the reaction chain that leads to uridine 5'-phosphate (Zrenner et al., 2006). The import transition added carbamoyl aspartate to the modules, which was produced by transition *E42* in the citrate module. *De novo* synthesis of uridine 5'-phosphate is highly energy consuming and in some tissues partly replaced by the recycling of already built compounds. Nevertheless, *de novo* synthesis of UMP is still needed to replenish the nucleotide stock (Moffatt and Ashihara, 2002). In the reduced network,

this complete synthesis pathway was merged into one transition, called $ctp(ctp(E42 + E43) + ctp(E44 + ctp(E61_119 + E61_80)))$. In the recursive reduction process, all of the reactions, forming the *de novo* synthesis of uridine 5'-phosphate, fulfilled the conditions necessary for a CTP reduction. As transition *E42* was part of the citrate module it could not be part of this MCTS. Nevertheless, the possible CTP reduction of the complete reaction chain, including *E42*, suggested a strong connection between the reactions of the uridine 5' phosphate *de novo* synthesis and could indicate an MCTS in the PN, covering all of them.

3.4.2. MCTS 2 and the Citrate Module

MCTS 2 consists of the transitions *E103*, *E104_f*, *E105*, and *E113_f*. This configuration occurred in the citrate module of the biologically driven decomposition as well as in the reduced network. Compared to the decompositions and the original network, the connections of the transitions have changed during the reduction process. In the subnets, the transitions form the reactions:

- $E103 : 79 \xrightarrow{E103} 4$
- $E104_f : 3 + 20 \xrightarrow{E104_f} 21 + 79$
- $E105 : 2 \xrightarrow{E105} 3$
- $E113_f : 9 + 21 \xrightarrow{E113_f} 20 + 2$

with 2 = glutamate, 3 = γ -aminobutyric acid (GABA), 4 = succinate, 9 = α -ketoglutarate, 20 = pyruvate, 21 = phosphoenolpyruvate, and 79 = succinate semialdehyde. In the reduced network, the metabolites 4 and 20 were combined by an ITP reduction, forming a new place. The connections of this new place were the combined connections of the merged original places, i.e., the new place connected to *E103* (as metabolite 4), *E104_f* (as metabolite 20), and *E113_f* (as metabolite 20) and led to the reduced reaction system

- $E103 : 79 \xrightarrow{E103} X$
- $E104_f : 3 + X \xrightarrow{E104_f} 21 + 79$
- $E105 : 2 \xrightarrow{E105} 3$
- $E113_f : 9 + 21E113_f \xrightarrow{E113_f} X + 2$

with 2 = glutamate, 3 = γ -aminobutyric acid (GABA), 9 = α -ketoglutarate, 21 = phosphoenolpyruvate, 79 = succinate semialdehyde, and *X* = the merged place (4 + 20). While the pathways for these transitions mentioned in the AraCyc database are glutamate degradation (*E10*, *E104_f*, and *E105*), and alanine degradation (*E113f*), other literature declare them as the GABA shunt (Bouché and Fromm, 2004). Finding an MCTS of these reactions strongly suggested a close interaction between them, indicating a possible network behavior consistent with the literature, because GABA is sufficient as sole nitrogen source for effective growth of *A. thaliana* (Breitkreuz et al., 1999), and the GABA shunt seems to play an important role in the reaction to oxidative stress (Bouché et al., 2003; Bouché and Fromm, 2004).

3.4.3. MCTS 3 and the Citrate Module

This MCTS constitutes of the transitions *E40*, *E133_f*, *E134*, *E135*, and *E136* in the citrate module. The transitions *E40*, *E133_f*, *E134*, and *E135* formed the urea cycle (Tischner et al., 2007), and transition *E136* modeled the degradation of urea (metabolite 25, Sirko and Brodzik, 2000). This MCTS is a collection of transitions forming and degrading urea. In the reduced network, two CTP reductions took place in this cycle. Initially, *E135* and *E40* were condensed to *ctp(E135 + E40)*, which was further flattened to *ctp(ctp(E135 + E40) + E136)* by the inclusion of *E136*. Together with *E134* and *E133_f*, this new reduced transition formed an MCTS in the reduced network. Urea is an important nitrogen source for plants (Polacco and Holland, 1993) and mainly believed to be predominantly synthesized by the urea cycle (Reinbothe and Mothes, 1962). This MCTS suggested the importance of the urea metabolism in the PN model of the core metabolism of *A. thaliana*.

4. CONCLUSION

In this paper, we presented a new semi-quantitative Petri net model of the metabolism of *A. thaliana* based on recent literature. Similar to a network of barley (Grafahrend-Belau et al., 2009), the model was manually developed and curated. To ensure the model's consistency, we used PN-based reduction and biologically motivated as well as graph-based decomposition and analysis techniques.

The final size of the complete PN model was 134 metabolites, 243 reactions, and 572 edges. The complexity of this model did not allow to compute all its transition invariants, which form the base for further analysis. To get a manageable set of TIs, we followed two strategies. First, we divided the model into four biology-driven subnetworks, the sucrose, the citrate, the UTP, and the shikimate subnetwork, and defined two modules, each consisting of two subnetworks. The sucrose module covers the sucrose and the UTP subnetwork, while the citrate module compiles the citrate and the shikimate subnetwork. The second strategy followed a graph-theoretic reduction of the model, applying common transition pairs and invariant transition pairs. Through the reduction, the network size decreased by ~50%. For all three subnetworks, we computed the TIs, easily showing that the subnetworks were CTI. To handle the amount of 27,646 TIs for the reduced model, we classified the TIs into trivial, INOUT, IN, OUT, and cyclic TIs. Because we could not discuss all the 27,646 TIs, we considered exemplarily one cyclic TI that describes a part of the starch synthesis and degradation and one TI that expresses a combination of parts of the nitrogen economy, the oxalacetate synthesis, the synthesis of 5-phosphoribosyl 1-pyrophosphate, and the synthesis of pyrimidines.

We demonstrated that the carbon fixation phase and the regeneration phase of the Calvin cycle strongly depends on each other. Additionally, potential steady-state pathways exist, which provided the fixed carbon to nearly all parts of the network, especially to the citric acid cycle. Moreover, the analysis showed a close cooperation of important metabolic pathways, e.g., the *de novo* synthesis of uridine-5-monophosphate, the γ -aminobutyric acid shunt, and the urea cycle.

The presented model provides a solid basis for further refinement, for example, by concentrations, gene expression data, and kinetic data for a quantitative analysis.

AUTHOR CONTRIBUTIONS

ES and IK initiated and supervised the project. JN has performed the network construction and analysis. All authors wrote the manuscript.

FUNDING

The project received funding from the Volkswagenstiftung to ES, from the LOEWE program Ubiquitin Networks (Ub-Net) of the State of Hesse (Germany) to IK and from the DFG Excellence Cluster CEF-Macromolecular Complexes to ES and IK.

ACKNOWLEDGMENTS

We would like to thank Oliver Mirus and Stefan Simm for guiding during the initial phase of the project and Jens Einloft for many fruitful discussions.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: <http://journal.frontiersin.org/article/10.3389/fgene.2017.00085/full#supplementary-material>

REFERENCES

- Ackermann, J., Einloft, J., Nöthen, J., and Koch, I. (2012). Reduction techniques for network validation in systems biology. *J. Theor. Biol.* 315, 71–80. doi: 10.1016/j.jtbi.2012.08.042
- Ackermann, J., and Koch, I. (2011). “Network decomposition in biology–computational challenges,” in *Proceedings of the Conference Models of Computation in Context. Computability in Europe*, vol. 7, eds H. Ganchev, B. Löwe, D. Normann, I. Soskov, and M. Soskova (Sofia: St. Kliment Ohridski University Press), 39–48.
- Acuña, V., Marchetti-Spaccamela, A., Sagot, M.-F., and Stougie, L. (2010). A note on the complexity of finding and enumerating elementary modes. *Biosystems* 99, 210–214. doi: 10.1016/j.biosystems.2009.11.004
- Alonso, J., Stepanova, A., Leisse, T., Kim, C., Chen, H., Shinn, P., et al. (2003). Genome-wide insertional mutagenesis of *Arabidopsis thaliana*. *Science* 301, 653–657. doi: 10.1126/science.1086391
- Arabidopsis-Genome-Initiative (2000). Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* 408, 796–815. doi: 10.1038/35048692
- Arnborg, S., Courcelle, B., Proskurowski, A., and Seese, D. (1993). An algebraic theory of graph reduction. *J. ACM* 40, 1134–1164. doi: 10.1145/174147.169807
- Aubert, S., Bligny, R., Douce, R., Gout, E., Ratcliffe, R., and Roberts, J. (2001). Contribution of glutamate dehydrogenase to mitochondrial glutamate metabolism studied by ^{13}C and ^{31}P nuclear magnetic resonance. *J. Exp. Bot.* 52, 37–45. doi: 10.1093/jxb/52.354.37
- Balazski, P., Lindauer, K., Einloft, J., Ackermann, J., and Koch, I. (2015). MonaLisa for stochastic simulations of Petri net models of biochemical systems. *BMC Bioinformatics* 16:215. doi: 10.1186/s12859-015-0596-y
- Bassham, J. A., Benson, A. A., Kay, L. D., Harris, A. Z., Wilson, A. T., and Calvin, M. (1954). The path of carbon in photosynthesis. XXI. The cyclic regeneration of carbon dioxide acceptor. *J. Amer. Chem. Soc.* 76, 1760–1770. doi: 10.1021/ja01636a012
- Baumgarten, B. (1996). *Petri Nets Basics and Applications (in German)*, 2nd Edn. Heidelberg: Berlin: Spektrum Akademischer Verlag.
- Berg, J. M., Tymoczko, J. L., and Stryer, L. (2002). *Biochemistry*, 5th Edn. New York, NY: W. H. Freeman.
- Berthelot, G. (1987). “Transformations and decompositions of nets,” in *Petri Nets: Central Models and Their Properties, Advances in Petri Nets, Lecture Notes in Computer Science*, eds W. Brauer, W. Reisig, and G. Rozenberg (Berlin: Heidelberg: Springer), 359–376.
- Billington, J., and Reisig, W. (1996). *Application and Theory of Petri Nets*. vol. 1091. Lecture notes Computer Science. Berlin: Heidelberg: Springer.
- Bortfeldt, R. H., Schuster, S., and Koch, I. (2011). Exhaustive analysis of the modular structure of the spliceosomal assembly network: a Petri net approach. *Stud. Health Technol. Inform.* 162, 244–278.
- Bouché, N., Fait, A., Bouchez, D., Möller, S. G., and Fromm, H. (2003). Mitochondrial succinic-semialdehyde dehydrogenase of the γ -aminobutyrate shunt is required to restrict levels of reactive oxygen intermediates in plants. *Proc. Natl. Acad. Sci. U.S.A.* 100, 6843–6848. doi: 10.1073/pnas.1037532100
- Bouché, N., and Fromm, H. (2004). GABA in plants: just a metabolite? *Trends Plant Sci.* 9, 110–115.
- Breitkreuz, K. E., Shelp, B. J., Fischer, W. N., Schwacke, R., and Rentsch, D. (1999). Identification and characterization of GABA, proline and quaternary ammonium compound transporters from *Arabidopsis thaliana*. *FEBS Lett.* 450, 280–284. doi: 10.1016/S0014-5793(99)00516-5
- Bruex, A., Kainkaryam, R. M., Wieckowski, Y., Kang, Y. H., Bernhardt, C., Xia, Y., et al. (2012). A gene regulatory network for root epidermis cell differentiation in *Arabidopsis*. *PLoS Genet.* 8:e1002446. doi: 10.1371/journal.pgen.1002446
- Buchanan, B. B., Gruissem, W., and Jones, R. L. (2000). *Biochemistry & Molecular Biology of Plants*. Rockville, MD: American Society of Plant Physiologists.
- Büchel, F., Rodriguez, N., Swainston, N., Wrzodek, C., Czauderna, T., Keller, K., et al. (2013). Path2Models: large-scale generation of computational models from biochemical pathway maps. *BMC Syst. Biol.* 7:116. doi: 10.1186/1752-0509-7-116
- Calvin, M. (1956). The photosynthetic carbon cycle. *J. Chem. Soc. (Resumed)* 1895–1915. doi: 10.1039/jr9560001895
- Cánovas, F. M., Avila, C., Cantón, F. R., Cañas, R. A., and de la Torre, F. (2007). Ammonium assimilation and amino acid metabolism in conifers. *J. Exp. Bot.* 58, 2307–2318. doi: 10.1093/jxb/erm051
- Chalot, M., Blaudez, D., and Brun, A. (2006). Ammonia: a candidate for nitrogen transfer at the mycorrhizal interface. *Trends Plant Sci.* 11, 263–266. doi: 10.1016/j.tplants.2006.04.005
- de Oliveira Dal’Molin, C. G., Quek, L.-E., Palfreyman, R. W., Brumby, S. M., and Nielsen, L. K. (2010). AraGEM, a genome-scale reconstruction of the primary metabolic network in *Arabidopsis*. *Plant Physiol.* 152, 579–589. doi: 10.1104/pp.109.148817
- Dersch, L. M., and Beckers, V. W. C. (2016). Green pathways: metabolic network analysis of plant systems. *Metab. Eng.* 34, 1–24. doi: 10.1016/j.ymben.2015.12.001
- Dey, P., and Harborne, J. (eds.). (1997). *Plant Biochem.* San Diego, CA: Academic Press.
- Eckardt, N. A. (2005). Photorespiration revisited. *Plant Cell Online* 17, 2139–2141. doi: 10.1105/tpc.105.035873
- Einloft, J., Ackermann, J., Nöthen, J., and Koch, I. (2013). MonaLisa - visualization and analysis of functional modules in biochemical networks. *Bioinformatics* 29, 1469–1470. doi: 10.1093/bioinformatics/btt165
- Fettke, J., Hejazi, M., Smirnova, J., Höchel, E., Stage, M., and Steup, M. (2009). Eukaryotic starch degradation: integration of plastidial and cytosolic pathways. *J. Exp. Bot.* 60, 2907–2922. doi: 10.1093/jxb/erp054
- Forde, B. G., and Lea, P. J. (2007). Glutamate in plants: metabolism, regulation, and signalling. *J. Exp. Bot.* 58, 2339–2358. doi: 10.1093/jxb/erm121
- Fulda, M., Shockey, J., Werber, M., Wolter, F. P., and Heinz, E. (2002). Two long-chain acyl-CoA synthetases from *Arabidopsis thaliana* involved in peroxisomal fatty acid β -oxidation. *Plant J.* 32, 93–103. doi: 10.1046/j.1365-3113X.2002.01405.x
- Gillespie, D. (1977). Exact stochastic simulation of coupled chemical reactions. *J. Phys. Chem.* 81, 2340–2361. doi: 10.1021/j100540a008
- Grafahrend-Belau, E., Schreiber, F., Heiner, M., Sackmann, A., Junker, B., Grunwald, S., et al. (2008). Modularization of biochemical networks based on classification of Petri net t-invariants. *BMC Bioinformatics* 9:90. doi: 10.1186/1471-2105-9-90
- Grafahrend-Belau, E., Schreiber, F., Koschützki, D., and Junker, B. H. (2009). Flux balance analysis of barley seeds: a computational approach to study systemic properties of central metabolism. *Plant Physiol.* 149, 585–598. doi: 10.1104/pp.108.129635
- Graindorge, M., Giustini, C., Jacomin, A. C., Kraut, A., Curien, G., and Matringe, M. (2010). Identification of a plant gene encoding glutamate/aspartate-prephenate aminotransferase: the last homeless enzyme of aromatic amino acids biosynthesis. *FEBS Lett.* 584, 4357–4360. doi: 10.1016/j.febslet.2010.09.037
- Grunwald, S., Speer, A., Ackermann, J., and Koch, I. (2008). Petri net modelling of gene regulation of the Duchenne muscular dystrophy. *Biosystems* 92, 189–205. doi: 10.1016/j.biosystems.2008.02.005
- Guimera, R., and Amaral, L. A. N. (2005). Functional cartography of complex metabolic networks. *Nature* 433, 895–900. doi: 10.1038/nature03288
- Guy, C., Kaplan, F., Kopka, J., Selbig, J., and Hincha, D. K. (2008). Metabolomics of temperature stress. *Physiol. Plant.* 132, 220–235.
- Heazlewood, J. L., Verboom, R. E., Tonti-Filippini, J., Small, I., and Millar, A. H. (2007). SUBA: the *Arabidopsis* subcellular database. *Nucl. Acids Res.* 35(Suppl 1), D213–D218. doi: 10.1093/nar/gkl863
- Heiner, M., and Koch, I. (2004). “Petri net based model validation in systems biology,” in *Applications and Theory of Petri Nets 2004, volume 3099 of Lecture Notes Computer Science*, eds J. Cortadella and W. Reisig (Berlin: Heidelberg: Springer), 216–237.
- Herrmann, K. M., and Weaver, L. M. (1999). The shikimate pathway. *Ann. Rev. Plant Biol.* 50, 473–503. doi: 10.1146/annurev.arplant.50.1.473
- Hickman, R., Hill, C., Penfold, C. A., Breeze, E., Bowden, L., Moore, J. D., et al. (2013). A local regulatory network around three NAC transcription factors in stress responses and senescence in *Arabidopsis* leaves. *Plant J.* 75, 26–39. doi: 10.1111/tjp.12194
- Humphreys, J. M., and Chapple, C. (2002). Rewriting the lignin roadmap. *Curr. Opin. Plant Biol.* 5, 224–229. doi: 10.1016/S1369-5266(02)00257-1
- Hur, M., Campbell, A., Almeida-de Macedo, M., Li, L., Ransom, N., Jose, A., et al. (2013). A global approach to analysis and interpretation of metabolic

- data for plant natural product discovery. *Nat. Prod. Rep.* 30, 565–583. doi: 10.1039/c3np20111b
- Kanehisa, M., Araki, M., Goto, S., Hattori, M., Hirakawa, M., Itoh, M., et al. (2008). KEGG for linking genomes to life and the environment. *Nucl. Acids Res.* 36(Suppl. 1), D480–D484. doi: 10.1093/nar/gkm882
- Koch, I., and Ackermann, J. (2013). On functional module detection in metabolic networks. *Metabolites* 3, 673–700. doi: 10.3390/metabo3030673
- Koch, I., and Heiner, M. (2008). “Petri nets,” in *Analysis of Biological Networks*, eds B. H. Junker and F. Schreiber (Wiley), 139–180.
- Koch, I., Junker, B. H., and Heiner, M. (2005). Application of Petri net theory for modelling and validation of the sucrose breakdown pathway in the potato tuber. *Bioinformatics* 21, 1219–1226. doi: 10.1093/bioinformatics/bti145
- Koch, I., Reisig, W., and Schreiber, F., editors (2011). *Modeling in Systems Biology: The Petri Net Approach*, volume 16 of Computational Biology, 1st Edn. London: Springer.
- Koornneef, M., and Meinke, D. (2010). The development of Arabidopsis as a model plant. *Plant J.* 61, 909–921. doi: 10.1111/j.1365-3113X.2009.04086.x
- Kossmann, J., and Lloyd, J. (2000). Understanding and influencing starch biochemistry. *Crit. Rev. Biochem. Mol. Biol.* 35, 141–196. doi: 10.1080/07352680091139204
- Kruger, N. J., and von Schaewen, A. (2003). The oxidative pentose phosphate pathway: structure and organisation. *Curr. Opin. Plant Biol.* 6, 236–246. doi: 10.1016/S1369-5266(03)00039-6
- Lautenbach, K. (1973). *Exakte Bedingungen der Lebendigkeit für eine Klasse von Petri-Netzen*. Number 82. St. Augustin: Gesellschaft für Mathematik und Datenverarbeitung.
- Lee-kwang, H., Favrel, J., and Baptiste, P. (1987). Generalized Petri net reduction method. *IEEE Trans. Syst.* 17, 297–303. doi: 10.1109/tsmc.1987.4309041
- Linka, M., and Weber, A. P. (2005). Shuffling ammonia between mitochondria and plastids during photorespiration. *Trends Plant Sci.* 10, 461–465. doi: 10.1016/j.tplants.2005.08.002
- Lu, Y., and Sharkey, T. (2006). The importance of maltose in transitory starch breakdown. *Plant Cell Environ.* 29, 353–366. doi: 10.1111/j.1365-3040.2005.01480.x
- Lucas, M. D., and Brady, S. M. (2013). Gene regulatory networks in the Arabidopsis root. *Curr. Opin. Plant Biol.* 16, 50–55. doi: 10.1016/j.pbi.2012.10.007
- Lunn, J. (2007). Compartmentation in plant metabolism. *J. Exp. Bot.* 58, 35–47. doi: 10.1093/jxb/erl134
- Masakapalli, S. K., Le Lay, P., Huddleston, J. E., Pollock, N. L., Kruger, N. J., and Ratcliffe, R. G. (2010). Subcellular flux analysis of central metabolism in a heterotrophic Arabidopsis cell suspension using steady-state stable isotope labeling. *Plant Physiol.* 152, 602–619. doi: 10.1104/pp.109.151316
- Matsuno, H., Doi, A., Nagasaki, M., and Miyano, S. (2000). “Hybrid Petri net representation of gene regulatory network,” in *Pacific Symposium Biocomputing*, vol. 87 (Singapore: World Scientific Press), 338–349.
- Meinke, D., Cherry, J., Dean, C., Rounsley, S., and Koornneef, M. (1998). *Arabidopsis thaliana*: a model plant for genome analysis. *Science* 282, 662–665. doi: 10.1126/science.282.5389.662
- Mintz-Oron, S., Meir, S., Malitsky, S., Rupp, E., Aharoni, A., and Shlomi, T. (2012). Reconstruction of Arabidopsis metabolic network models accounting for subcellular compartmentalization and tissue-specificity. *Proc. Natl. Acad. Sci. U.S.A.* 109, 339–344. doi: 10.1073/pnas.1100358109
- Moffatt, B. A., and Ashihara, H. (2002). Purine and pyrimidine nucleotide synthesis and metabolism. *Arabidopsis Book/Amer. Soc. Plant Biol.* 1:e0018. doi: 10.1199/tab.0018
- Mueller, L. A., Zhang, P., and Rhee, S. Y. (2003). AraCyc: a biochemical pathway database for Arabidopsis. *Plant Physiol.* 132, 453–460. doi: 10.1104/pp.102.017236
- Murata, T. (1989). Petri nets: properties, analysis and applications. *Proc. IEEE* 77, 541–580. doi: 10.1109/5.24143
- Nöthen, J. (2009). *Metabolische Netzwerke in Pflanzen*. Johann Wolfgang Goethe-University Frankfurt am Main, Diploma Thesis.
- Nöthen, J. (2014). *Mathematical Modeling of Arabidopsis thaliana with Focus on Network Decomposition and Reduction*. Johann Wolfgang Goethe-University Frankfurt am Main, Ph.D. thesis.
- Parry, M. A. J., Andralojc, P. J., Mitchell, R. A. C., Madgwick, P. J., and Keys, A. J. (2003). Manipulation of Rubisco: the amount, activity, function and regulation. *J. Exp. Bot.* 54, 1321–1333. doi: 10.1093/jxb/erg141
- Peltier, J.-B., Cai, Y., Sun, Q., Zabrouskov, V., Giacomelli, L., Rudella, A., et al. (2006). The oligomeric stromal proteome of *Arabidopsis thaliana* chloroplasts. *Mol. Cell. Proteomics* 5, 114–133. doi: 10.1074/mcp.M500180-MCP200
- Petri, C. A. (1962). *Communication with Automats*. German. Ph.D. Thesis, Technische Hochschule Darmstadt, in German.
- Poolman, M. G., Miguet, L., Sweetlove, L. J., and Fell, D. A. (2009). A genome-scale metabolic model of Arabidopsis and some of its properties. *Plant Physiol.* 151, 1570–1581. doi: 10.1104/pp.109.141267
- Radrich, K., Tsuruoka, Y., Dobson, P., Gevorgyan, A., Swainston, N., Baart, G., et al. (2010). Integration of metabolic databases for the reconstruction of genome-scale metabolic networks. *BMC Syst. Biol.* 4:114. doi: 10.1186/1752-0509-4-114
- Raines, C. (2003). The Calvin cycle revisited. *Photosynthesis Res.* 75, 1–10. doi: 10.1023/A:1022421515027
- Raman, S. B., and Rathinasabapathi, B. (2004). Pantothenate synthesis in plants. *Plant Sci.* 167, 961–968. doi: 10.1016/j.plantsci.2004.06.019
- Ravasz, E., Somera, A. L., Mongru, D. A., Oltvai, Z. N., and Barabási, A.-L. (2002). Hierarchical organization of modularity in metabolic networks. *Science* 297, 1551–1555. doi: 10.1126/science.1073374
- Reddy, V. N., Mavrovouniotis, M. L., and Liebman, M. N. (1993). “Petri net representations in metabolic pathways,” in *Proceedings ISMB*, eds L. Hunter, D. B. Searls, and J. W. Shavlik (Menlo Park, CA: AAAI Press), 328–336.
- Reinbothe, H., and Mothes, K. (1962). Urea, ureides, and guanidines in plants. *Annu. Rev. Plant Physiol.* 13, 129–149. doi: 10.1146/annurev.pp.13.060162.001021
- Reiter, W. (2008). Biochemical genetics of nucleotide sugar interconversion reactions. *Curr. Opin. Plant Biol.* 11, 236–243. doi: 10.1016/j.pbi.2008.03.009
- Reumann, S., Ma, C., Lemke, S., and Babujee, L. (2004). AraPeroX: a database of putative Arabidopsis proteins from plant peroxisomes. *Plant Physiol.* 136, 2587–2608. doi: 10.1104/pp.104.043695
- Roy, H., and Andrews, T. (2004). “Rubisco: assembly and mechanism,” in *Photosynthesis, volume 9 of Advances in Photosynthesis and Respiration*, eds R. C. Leegood, T. D. Sharkey, and S. Caemmerer (Dordrecht: Springer), 53–83.
- Sackmann, A., Heiner, M., and Koch, I. (2006). Application of Petri net based analysis techniques to signal transduction pathways. *BMC Bioinformatics* 7:482. doi: 10.1186/1471-2105-7-482
- Saha, R., Suthers, P. F., and Maranas, C. D. (2011). *Zea mays* iRS1563: a comprehensive genome-scale metabolic reconstruction of maize metabolism. *PLoS ONE* 6:e21784. doi: 10.1371/journal.pone.0021784
- Schrijver, A. (1998). *Theory of Linear and Integer Programming*. Weinheim: Wiley-VCH.
- Schuster, S., and Hilgetag, C. (1994). On elementary flux modes in biochemical reaction systems at steady state. *J. Biol. Syst.* 2, 165–182. doi: 10.1142/S0218339094000131
- Schuster, S., Hilgetag, C., Woods, J., and Fell, D. (2002). Reaction routes in biochemical reaction systems: algebraic properties, validated calculation procedure and example from nucleotide metabolism. *J. Math. Biol.* 45, 153–181. doi: 10.1007/s002850200143
- Sessions, A., Burke, E., Presting, G., George Aux, G., McElver, J., Patton, D., et al. (2002). A high-throughput Arabidopsis reverse genetics system. *Plant Cell* 14, 2985–2994. doi: 10.1105/tpc.004630
- Sirko, A., and Brodzik, R. (2000). Plant ureases: roles and regulation. *Acta Biochim. Polonica* 47, 1189–1115.
- Streb, S., and Zeeman, S. C. (2012). Starch metabolism in Arabidopsis. *Arabidopsis Book/Amer. Soc. Plant Biol.* 10:e0160. doi: 10.1199/tab.0160
- Szydlowski, N., Ragel, P., Raynaud, S., Lucas, M. M., Roldán, I., Montero, M., et al. (2009). Starch granule initiation in Arabidopsis requires the presence of either class IV or class III starch synthases. *Plant Cell Online* 21, 2443–2457. doi: 10.1105/tpc.109.066522
- Temple, S. J., Vance, C. P., and Gantt, J. S. (1998). Glutamate synthase and nitrogen assimilation. *Trends Plant Sci.* 3, 51–56.
- Tischner, R., Galli, M., Heimer, Y. M., Bielefeld, S., Okamoto, M., Mack, A., et al. (2007). Interference with the citrulline-based nitric oxide synthase assay by argininosuccinate lyase activity in Arabidopsis extracts. *FEBS J.* 274, 4238–4245. doi: 10.1111/j.1742-4658.2007.05950.x
- Uzam, M. (2004). The use of the Petri net reduction approach for an optimal deadlock prevention policy for flexible manufacturing systems. *Int. J. Adv. Manuf. Technol.* 23, 204–219. doi: 10.1007/s00170-002-1526-5

- Van Norman, J. M., and Benfey, P. N. (2009). *Arabidopsis thaliana* as a model organism in systems biology. *Wiley Interdiscip. Rev. Syst. Biol. Med.* 1, 372–379. doi: 10.1002/wsbm.25
- Vasilevski, A., Giorgi, F. M., Bertinetti, L., and Usadel, B. (2012). LASSO modeling of the *Arabidopsis thaliana* seed/seedling transcriptome: a model case for detection of novel mucilage and pectin metabolism genes. *Mol. Biosyst.* 8, 2566–2574. doi: 10.1039/c2mb25096a
- Wilkie, S. E., and Warren, M. J. (1998). Recombinant expression, purification, and characterization of three isoenzymes of aspartate aminotransferase from *Arabidopsis thaliana*. *Protein Exp. Purif.* 12, 381–389. doi: 10.1006/prep.1997.0845
- Williams, T. C., Miguët, L., Masakapalli, S. K., Kruger, N. J., Sweetlove, L. J., and Ratcliffe, R. G. (2008). Metabolic network fluxes in heterotrophic *Arabidopsis* cells: stability of the flux distribution under different oxygenation conditions. *Plant Physiol.* 148, 704–718. doi: 10.1104/pp.108.125195
- Zhang, P., Foerster, H., Tissier, C. P., Mueller, L., Paley, S., Karp, P. D., et al. (2005). MetaCyc and AraCyc. Metabolic pathway databases for plant research. *Plant Physiol.* 138, 27–37. doi: 10.1104/pp.105.060376
- Zrenner, R., Stitt, M., Sonnewald, U., and Boldt, R. (2006). Pyrimidine and purine biosynthesis and degradation in plants. *Ann. Rev. Plant Biol.* 57, 805–836. doi: 10.1146/annurev.arplant.57.032905.105421

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

The reviewer GR and handling editor declared their shared affiliation, and the handling editor states that the process nevertheless met the standards of a fair and objective review.

Copyright © 2017 Koch, Nöthen and Schleiff. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



Genetic Network Inference Using Hierarchical Structure

Shuhei Kimura^{1*}, Masato Tokuhisa¹ and Mariko Okada-Hatakeyama²

¹ Department of Information and Electronics, Graduate School of Engineering, Tottori University, Tottori, Japan, ² Laboratory for Integrated Cellular Systems, RIKEN Center for Integrative Medical Sciences, Yokohama, Japan

OPEN ACCESS

Edited by:

Rui Alves,
Universitat de Lleida, Spain

Reviewed by:

Ovidiu Radulescu,
Université Montpellier 2, France
Daisuke Tominaga,
National Institute of Advanced
Industrial Science and Technology,
Japan

*Correspondence:

Shuhei Kimura
kimura@eeecs.tottori-u.ac.jp

Specialty section:

This article was submitted to
Systems Biology,
a section of the journal
Frontiers in Physiology

Received: 02 November 2015

Accepted: 05 February 2016

Published: 23 February 2016

Citation:

Kimura S, Tokuhisa M and
Okada-Hatakeyama M (2016) Genetic
Network Inference Using Hierarchical
Structure. *Front. Physiol.* 7:57.
doi: 10.3389/fphys.2016.00057

Many methods for inferring genetic networks have been proposed, but the regulations they infer often include false-positives. Several researchers have attempted to reduce these erroneous regulations by proposing the use of a priori knowledge about the properties of genetic networks such as their sparseness, scale-free structure, and so on. This study focuses on another piece of a priori knowledge, namely, that biochemical networks exhibit hierarchical structures. Based on this idea, we propose an inference approach that uses the hierarchical structure in a target genetic network. To obtain a reasonable hierarchical structure, the first step of the proposed approach is to infer multiple genetic networks from the observed gene expression data. We take this step using an existing method that combines a genetic network inference method with a bootstrap method. The next step is to extract a hierarchical structure from the inferred networks that is consistent with most of the networks. Third, we use the hierarchical structure obtained to assign confidence values to all candidate regulations. Numerical experiments are also performed to demonstrate the effectiveness of using the hierarchical structure in the genetic network inference. The improvement accomplished by the use of the hierarchical structure is small. However, the hierarchical structure could be used to improve the performances of many existing inference methods.

Keywords: genetic network, hierarchical random graph, hierarchical structure, bootstrap method, simulated annealing

1. INTRODUCTION

A genetic network is a functioning circuit in living cells at the gene level. From one viewpoint, a genetic network can be seen as an abstract mapping of an actual biochemical network consisting of genes, proteins, metabolites, and so on. The analysis of genetic networks is conceived as one of the promising ways to understand biological systems. The mathematical modeling of genetic networks has therefore become an important theme in systems biology.

Many studies have sought to develop computational methods for inferring genetic networks from observed gene expression patterns (Larrañaga et al., 2006; Chou and Voit, 2009; Hecker et al., 2009). Often, however, these methods infer false-positive regulations along with true-positive regulations. These erroneous regulations must be decreased if we are to successfully analyze the inferred genetic networks. One possible approach to remove these erroneous regulations from the inferred genetic networks is to use a priori knowledge about the networks. Several researchers have introduced a priori knowledge about the properties of genetic networks, such as their sparseness, scale-free structure, and so on, into methods for inferring genetic networks (see, e.g., Kikuchi et al., 2003; Daisuke and Horton, 2006).

This study focuses on another type of a priori knowledge, namely, that biochemical networks exhibit hierarchical structures (Clauset et al., 2008). The hierarchical structure in a network is a property having vertices that cluster together in groups, which then join to form groups of groups, and so forth, from the lowest levels of organization up to the level of the entire network. If we know the hierarchical structure in the target genetic network, we can improve a genetic network inferred by an inference method. That is, we can conclude that the regulations inferred by the method are unreasonable if they are inconsistent with the hierarchical structure. The hierarchical structure in a given network can be detected using a method based on the hierarchical random graph model (Clauset et al., 2008). While this detection method assumes that the erroneous regulations in a given network are infrequent, erroneous regulations actually tend to be abundant in a network inferred by a method for inferring genetic networks. Even if we simply used Clauset's method for the analysis of a genetic network, a reasonable hierarchical structure would be difficult to obtain.

In order to detect a hierarchical structure correctly, this study first infers multiple genetic networks from the observed gene expression data using a genetic network inference method in combination with a bootstrap method (Efron, 1979). We then extract a hierarchical structure from the inferred genetic networks that is consistent with most of the networks. As some erroneous regulations seem to be rarely inferred by the bootstrap method, we speculated that the proposed approach could reduce the effect of these erroneous regulations on the hierarchical structure detection. In this study, we extract a hierarchical structure from multiple genetic networks using the detection method proposed by Clauset et al. (2008) with modifications and then use the hierarchical structure obtained to assess the confidence values of the regulations. Through numerical experiments, we then demonstrate the effectiveness of the use of the hierarchical structure in the genetic network inference.

2. DETECTING HIERARCHICAL STRUCTURES

2.1. Hierarchical Random Graph Model

Clauset et al. (2008) have proposed a method for detecting a hierarchical structure in a given network. Their method describes the given network as an undirected graph where the vertices and edges represent genes and interactions between them, respectively, in the genetic network inference. Note therefore that, while the method for inferring genetic networks generally treats a genetic network as a directed graph, the method for detecting hierarchical structures must treat it as an undirected graph.

The method proposed by Clauset et al. (2008) uses a hierarchical random graph model $H(D, \theta)$ to represent a hierarchical structure of a network consisting of N vertices, where D is a rooted binary tree having N leaf nodes and $N - 1$ internal nodes, and $\theta = (\theta_1, \theta_2, \dots, \theta_{N-1})$ (see **Figure 1**). Each of the N leaf nodes of D corresponds to each of the vertices of the given

network. The $N - 1$ internal nodes, which we represent here as D_1, D_2, \dots, D_{N-1} , indicate the hierarchical relationship among the vertices of the given network. Note that each pair of vertices in the given network has a unique internal node in D as their lowest common ancestor. The internal node D_i has a parameter θ_i . The parameter θ_i represents the probability that the given network has an edge between vertices where D_i is the lowest common ancestor in D . When the vertices u and v have the internal node D_i as their lowest common ancestor, therefore, it means that the network has an edge between these vertices with the probability θ_i . The model $H(D, \theta)$ has an ability to capture the hierarchical structure of the given network. On the other hand, $H(D, \theta)$ is also conceived as a generative model that allows us to generate artificial networks with a specified hierarchical structure.

The method proposed by Clauset et al. (2008) tries to find D and θ of $H(D, \theta)$, a model that serves well in representing the hierarchical structure of the given single network. The method proposed in this study, on the other hand, searches for them using multiple genetic networks inferred by the bootstrap approach.

2.2. Problem Definition

The method proposed by Clauset et al. (2008) uses the maximum likelihood estimation for the hierarchical structure detection. Similarly, the method we propose here uses the maximum likelihood estimation to extract a hierarchical structure from the given networks. Here, therefore, we obtain the rooted binary tree D and the parameter vector θ by maximizing a probability that the given networks are generated from the model $H(D, \theta)$. The detection of the hierarchical structure in this study is thus defined as a maximization problem of the log-likelihood function (Kimura and Okada-Hatakeyama, 2015)

$$\log L(D, \theta) = \sum_{j=1}^{N_g} \sum_{i=1}^{N-1} \left[E_i^j \log \theta_i + (L_i R_i - E_i^j) \log(1 - \theta_i) \right], \quad (1)$$

where N_g is the number of the given networks, N is the number of vertices contained in each network, and E_i^j is the number of edges in the j -th network between vertices having D_i as the lowest common ancestor in D . L_i and R_i are the number of leaf nodes of the left and right subtrees, respectively, rooted at D_i .

From the optimality conditions on the maximization problem of the function (Equation 1), i.e., $\frac{\partial \log L}{\partial \theta_i} = 0$, ($i = 1, 2, \dots, N - 1$), we obtain

$$\theta_i = \frac{\sum_{j=1}^{N_g} E_i^j}{N_g L_i R_i}, \quad (i = 1, 2, \dots, N - 1) \quad (2)$$

The equations above indicate that the appropriate values for the parameters θ_i 's are easily obtained for a given binary tree D . Our method thus extracts the hierarchical structure only by searching for the optimal D , as described below.

2.3. Optimization Algorithm

The method proposed by Clauset et al. (2008) extracts a hierarchical structure from only a single network. The given data are insufficient, so many hierarchical random graph models seem

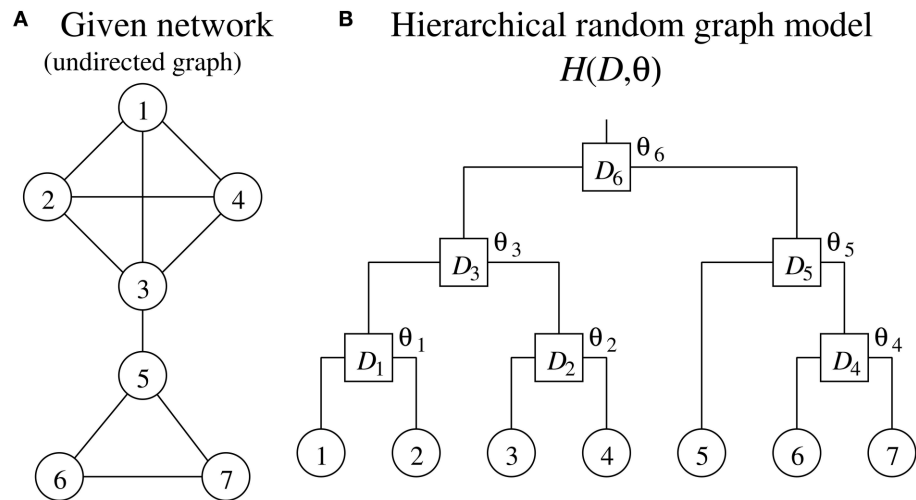


FIGURE 1 | (A) A network and (B) the corresponding hierarchical random graph model.

to match the given network well. Their method thus generates multiple models using a Markov chain Monte Carlo method (Chib and Greenberg, 1995), and then averages them to obtain the hierarchical structure.

In our study, the use of multiple networks to detect the hierarchical structure allows us to search for a single optimum model using a simulated annealing (Kirkpatrick et al., 1983). Our method thus optimizes the objective function (Equation 1) according to the following procedure.

[Algorithm for maximizing the function (Equation 1)]

1. Randomly generate a rooted binary tree, *BestTree*, with N leaf nodes, where N is the number of vertices contained in each of the given networks. Note that each leaf node of the tree corresponds to each vertex of the given networks. Compute the objective value of *BestTree* using the function (Equation 1). To compute this value for the function (Equation 1), we must first give the parameters θ_i along with a binary tree D . As mentioned in the section Problem Definition, however, this study directly computes the values for θ_i according to the Equation (2). Set T to T_{start} .
2. Copy *BestTree* to *CurrentTree*.
3. Set *Counter* to 0.
4. Copy *CurrentTree* to *TestTree*.
5. Select an internal node of *TestTree* randomly. Then, modify the structure of *TestTree* by applying 'Exchange' or 'Rotate' randomly to the selected node. "Exchange" and "Rotate" are operators that alter the structure of the subtree rooted at the selected node, as shown in Figure 2. After the modification, compute the objective value of *TestTree*.
6. If the objective value of *TestTree* is better than that of *BestTree*, copy *TestTree* to *BestTree*.
7. Copy *TestTree* to *CurrentTree* with a probability

$$\min \left\{ 1, \exp \left(-\frac{Obj_c - Obj_t}{T} \right) \right\},$$

where Obj_c and Obj_t are the objective values of *CurrentTree* and *TestTree*, respectively.

8. $Counter \leftarrow Counter + 1$.
9. Return to the step 4 if $Counter < N_{max}$.
10. $T \leftarrow \gamma T$.
11. Return to the step 2 if $T > T_{end}$. Otherwise, output *BestTree* and stop.

T_{start} , T_{end} , N_{max} , and γ in the algorithm above are constant parameters. For this study, we set their values to 1000, 0.1, $1000N$, and 0.99, respectively.

3. ASSIGNMENT OF CONFIDENCE VALUES TO REGULATIONS

As mentioned previously, this study first infers N_g genetic networks from the observed time-series of the gene expression levels. Any inference method capable of producing multiple genetic networks will serve this purpose. Here, however, we decided to use a method proposed by Kimura et al. (2010) for the generation of multiple genetic networks within a relatively short computation time by combining the LPM-based inference method (Kimura et al., 2009a) with the bootstrap method. We refer to this inference method as the BS-LPM inference method.

While the BS-LPM inference method distinguishes the regulation of the n -th gene from the m -th gene and vice versa, the method for detecting hierarchical structures described in the section Detecting Hierarchical Structures makes no such distinction. Here, therefore, we take the following step to transform the inferred genetic networks to the networks for our method for detecting the hierarchical structure: when the j -th genetic network inferred by the BS-LPM inference method contains the regulation of the n -th gene from the m -th gene, the regulation of the m -th gene from the n -th gene, or both, we add an edge between the n -th and m -th vertices to the j -th network for our detection method. The BS-LPM inference method is

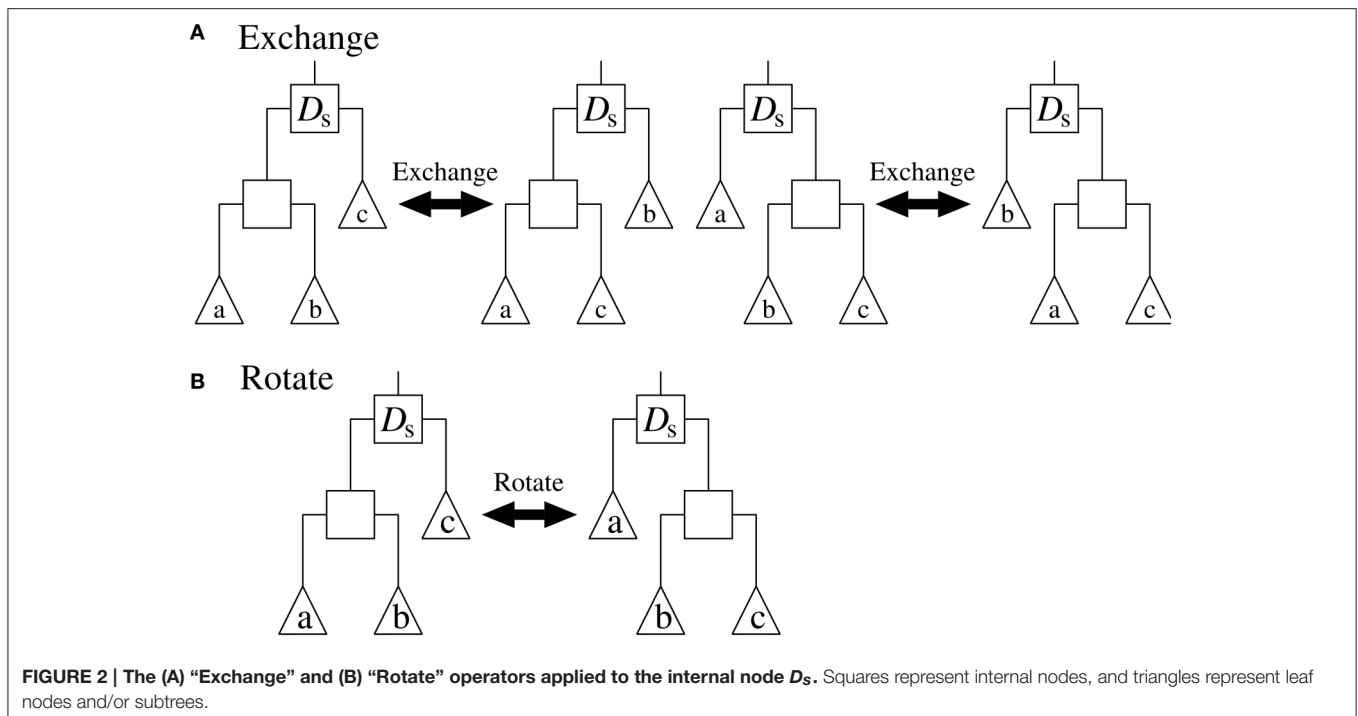


FIGURE 2 | The (A) “Exchange” and (B) “Rotate” operators applied to the internal node D_s . Squares represent internal nodes, and triangles represent leaf nodes and/or subtrees.

also capable of inferring an auto-regulation/auto-degradation, i.e., a regulation of a gene by itself. Here, however, we have to remove auto-regulations/auto-degradations from the networks, as our detection method cannot cope with them. Inferred networks usually contain auto-regulations/auto-degradations, because inference methods often infer the degradation of transcripts of a gene as a regulation of the gene by itself. We would not always need to search for regulations that usually exist. As such, the inference of auto-regulations/auto-degradations is not always essential for the inference of actual genetic networks. In order to detect a hierarchical structure in our target network, we next apply our detection method to the networks transformed above.

The confidence values of regulations can be evaluated solely based on the probabilities that the genetic networks inferred by the BS-LPM inference method contain the regulations. The hierarchical random graph model $H(D, \theta)$ in our method provides the probabilities that the target network has interactions between genes, which enable us to assign the confidence values to regulations on that basis, as well. We therefore try to improve the confidence values of regulations in this study by combining the probabilities evaluated by the BS-LPM inference method with those evaluated by $H(D, \theta)$. This study simply computes the combined confidence value of the regulation of the n -th gene from the m -th gene, $p_{n,m}$, by

$$p_{n,m} = \eta p_{n,m}^B + (1 - \eta) p_{n,m}^H, \quad (3)$$

where η ($0 \leq \eta \leq 1$) is a constant parameter, and $p_{n,m}^B$ and $p_{n,m}^H$ are the probabilities assigned to the regulation of the n -th gene from the m -th gene evaluated by the BS-LPM inference method and $H(D, \theta)$, respectively. Note here that $H(D, \theta)$ disregards the

directions of regulations. While the values for $p_{n,m}^B$ and $p_{m,n}^B$ are basically different from each other, therefore, $p_{n,m}^H$ and $p_{m,n}^H$ always have the same value.

Note that the hierarchical random graph model $H(D, \theta)$ is extracted from the networks inferred by the BS-LPM inference method. Therefore, we should not depend too much on the results obtained from $H(D, \theta)$. In this study, thus, we mainly use the extracted hierarchical structure to rank the regulations that are assigned the same probability value by the BS-LPM inference method. For this purpose, this study sets the parameter η to $1 - \frac{1}{N_g}$.

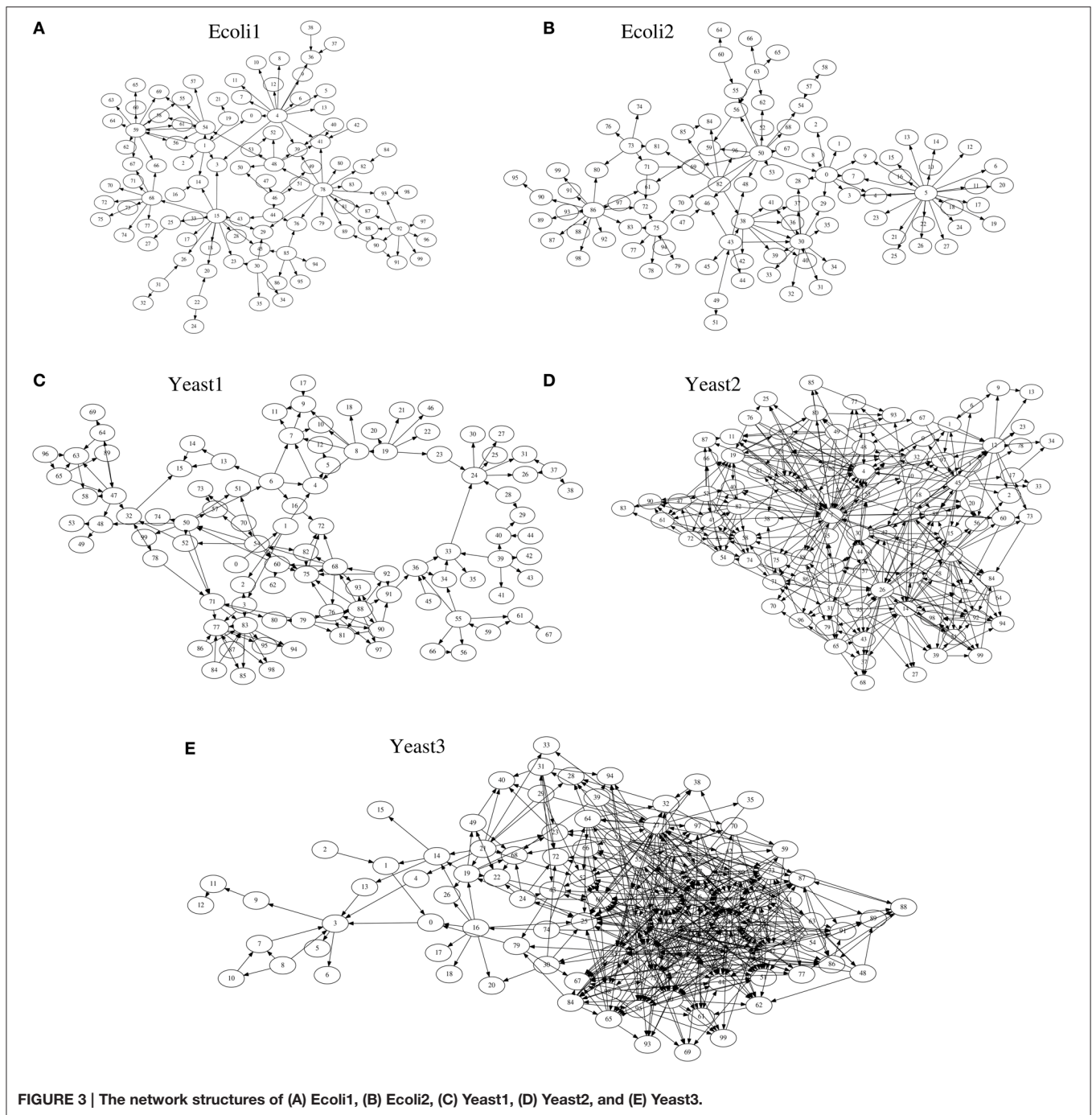
4. NUMERICAL EXPERIMENTS

4.1. Analysis of DREAM3 Networks

From here, we will describe a series of experiments performed with five artificial genetic networks to check whether or not the use of the hierarchical structure is efficient for the inference of genetic networks.

4.1.1. Experimental Setup

As target networks, we used a series of S-system models (Voit, 2000) consisting of 100 genes ($N = 100$), with topologies identical to those of the five networks provided by the DREAM3 *in silico* network challenges, i.e., Ecoli1, Ecoli2, Yeast1, Yeast2, and Yeast3 (<http://dreamchallenges.org/>) (Figure 3). The DREAM3 networks have often been used to check the performance of genetic network inference methods (see e.g., Lim et al., 2013). The design of these networks is based on actual biochemical networks and therefore reflects the actual topological properties. Note here that our method



for detecting hierarchical structures only uses the topological properties of genetic networks inferred by a genetic network inference method. Although the target networks are artificial, the experiments we describe here could confirm the effectiveness of the use of the hierarchical structure for the genetic network inference. DREAM3, on the other hand, describes these networks using a model different from the S-system model (Prill et al., 2010). While the model used in DREAM3 considers the effect of the intrinsic noise, the S-system model disregards

it. The BS-LPM inference method used in this study also disregards the intrinsic noise, so we used the S-system model to describe the target networks. Note that the purpose of the experiments here was not to assess the performance of the inference method but to check the effectiveness of the use of the hierarchical structure for the genetic network inference. We could therefore demonstrate the effectiveness of the use of the hierarchical structure even when using the S-system model.

The S-system model is a set of differential equations of the form

$$\frac{dX_n}{dt} = \alpha_n \prod_{m=1}^N X_m^{g_{n,m}} - \beta_n \prod_{m=1}^N X_m^{h_{n,m}} \quad (n = 1, 2, \dots, N), \quad (4)$$

where X_n is the n -th state variable, N is the number of components in the network, and $\alpha_n (> 0)$, $\beta_n (> 0)$, $g_{n,m}$, and $h_{n,m}$ are model parameters. In the genetic network inference, X_n is the expression level of the n -th gene and N is the number of genes contained in the target network. As the parameters $g_{n,m}$'s and $h_{n,m}$'s determine the topology of the network, we constructed the target networks by changing their values. In instances where the original DREAM3 network has the regulation of the n -th gene from the m -th gene, we chose a value for $g_{n,m}$ randomly from $[-1, -0.5] \cup [0.5, 1]$. Otherwise, $g_{n,m}$ was set to 0.0. The parameter $h_{n,n}$ was set to 1.0 in order to simulate the auto-degradation, and the other $h_{n,m}$'s ($n \neq m$) were set to 0.0. The parameters α_n and β_n were all set to 1.0. We determined to use this parameter setting based on the reference (Kimura et al., 2009a). The numbers of regulations contained in Ecoli1, Ecoli2, Yeast1, Yeast2, and Yeast3, excluding auto-degradations, were 125, 119, 166, 389, and 551, respectively. As the inference ability of the proposed approach might depend on the values for the model parameters, we changed the random parameter values in every trial. Ten trials were performed on each of the five target networks.

As the observed gene expression patterns, 100 sets of time-series data, each covering 100 genes, were computed from the differential Equations (4) on each of the target models. The sets began from randomly generated initial values in $[0.0, 2.0]$, and 11 observations with 0.4 time intervals between two adjacent observations were assigned to each gene in each set. In a practical application, these sets would be obtained by actual biological experiments under different experimental conditions. The measurement noise was simulated by adding 10% Gaussian noise to the computed time-series data. By applying the BS-LPM inference method (Kimura et al., 2010) to the generated gene expression data, we inferred 100 networks ($N_g = 100$). We used the recommended values for the parameters of the BS-LPM inference method, namely, $\sigma = 0.15$, $C_1 = \frac{200}{N\sqrt{K}}$, $C_2 = 0.4C_1$, and $\delta = 0.05$, where N is the number of genes contained in the target network and K is the number of measurements. Thus, $N = 100$ and $K = 100 \times 11 = 1100$ in these experiments.

In order to obtain a hierarchical random graph model $H(D, \theta)$, we then applied the hierarchical structure detection method described in the section Detecting Hierarchical Structures to the N_g generated genetic networks. We then used the hierarchical random graph model obtained to compute the confidence values of the regulations, as described in the section Assignment of Confidence Values to Regulations. The constant parameter for computing the confidence values, η , was set to $1 - \frac{1}{N_g} = 0.99$. As mentioned previously, we mainly uses the extracted hierarchical structure to rank the regulations that are assigned the same confidence value by the BS-LPM inference

method. This study therefore did not depend too much on the hierarchical structure $H(D, \theta)$.

4.1.2. Results

As described previously, the proposed approach and the BS-LPM inference method were both capable of assigning the confidence values to all of the candidate regulations. In this study, we checked the performance of these methods by constructing a network of regulations whose confidence values exceeded a threshold and then comparing it with the target network. We checked the performance using the recall and the precision. The recall and the precision are defined as

$$\text{recall} = \frac{TP}{TP + FN}, \quad \text{precision} = \frac{TP}{TP + FP},$$

where TP , FP , and FN are the numbers of true-positive, false-positive, and false-negative regulations, respectively. Note that we transformed the genetic networks inferred by the BS-LPM inference method into undirected graphs for detecting their hierarchical structure. When evaluating the performance, however, we distinguished the regulation of the n -th gene from the m -th gene and vice versa, i.e., we treated the networks as directed graphs. We also disregarded auto-regulations/auto-degradations in the evaluation.

Figure 4 shows samples of the recall-precision curves obtained by the proposed approach and by the BS-LPM inference method by changing the threshold for the confidence value. We previously described how closely our method depends on the BS-LPM inference method. As the figure shows, the performance of our approach was therefore similar to that of the BS-LPM inference method. Meanwhile, the figure also shows that the use of the hierarchical structure improved the precision of our approach. This higher precision is a preferable feature, since biologists must experimentally validate the inferred regulations in actual applications. The BS-LPM inference method required about 4.12 h on a personal computer (Core i5-4670) to obtain $N_g (= 100)$ genetic networks from the given gene expression patterns. The hierarchical structure detection method described in the section Detecting Hierarchical Structures required about 2.91 h on the same computer to extract a hierarchical structure from the generated genetic networks.

We quantified the performance of the proposed approach and the BS-LPM inference method in this study using the area under the recall-precision curve (AURPC). **Table 1** lists the averaged AURPCs of the two methods on the problems of Ecoli1, Ecoli2, Yeast1, Yeast2, and Yeast3. Our approach outperformed the BS-LPM inference method on most of the $5 \times 10 = 50$ trials with respect to the AURPC, but its performance was still inferior in seven of the trials. The inferior performance in those seven failed trials was presumably due to a failure of our approach to detect the hierarchical structures in the target networks. Four of the failed trials were performed on the Ecoli2 problem and the other three were performed on Yeast1, Yeast2, and Yeast3. As shown in **Figure 3B**, a number of genes in Ecoli2 are regulated by only single genes. Our approach failed to adequately analyze networks with this property, as some regulations erroneously inferred by the BS-LPM inference method easily caused the formation of

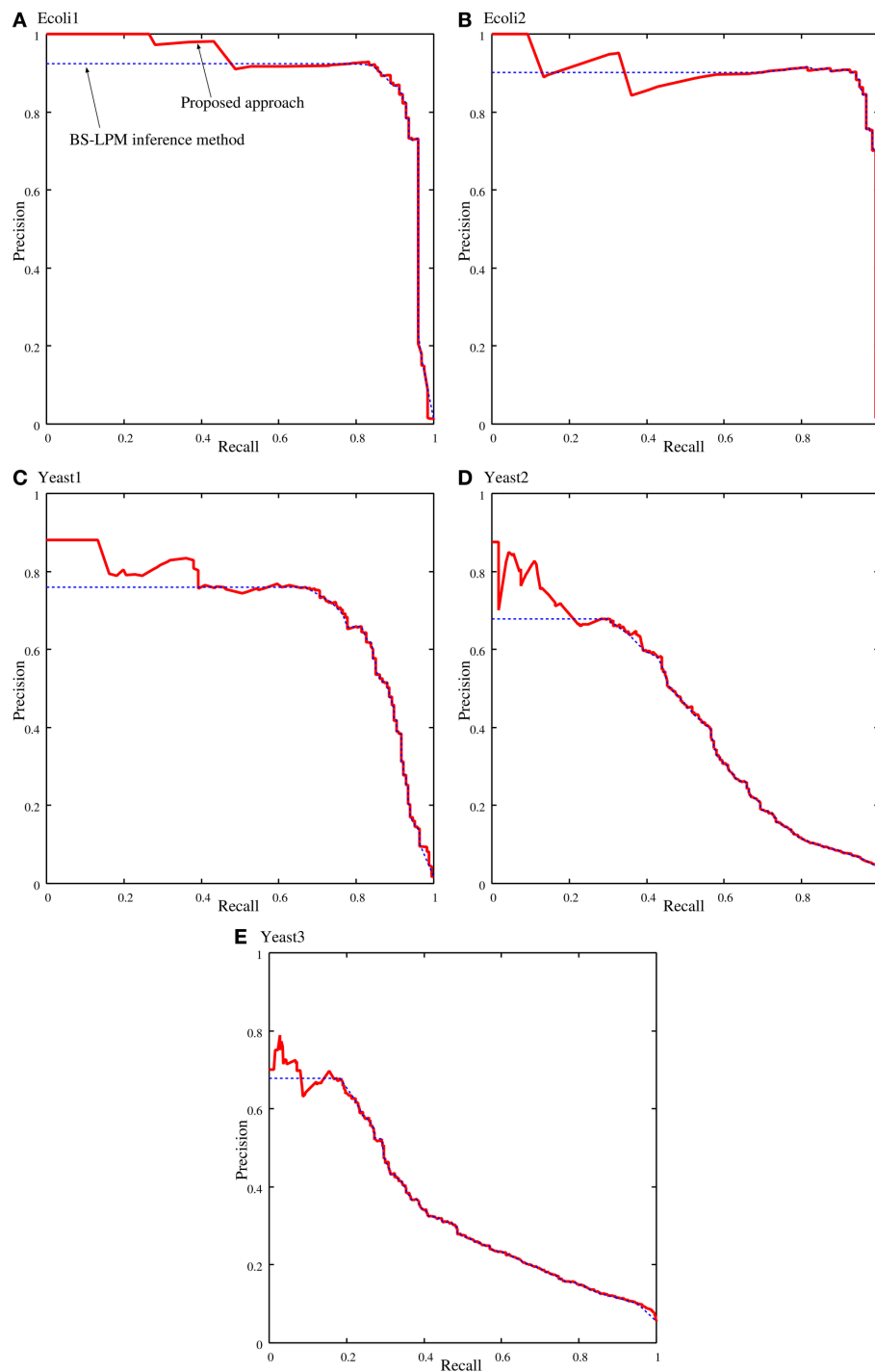


FIGURE 4 | Samples of the recall-precision curves of the proposed approach and the BS-LPM inference method on the problems of (A) Ecoli1, (B) Ecoli2, (C) Yeast1, (D) Yeast2, and (E) Yeast3. Solid and dotted lines represent the performance of the proposed approach and the BS-LPM inference method, respectively.

erroneous gene clusters. Ecoli2 would model a network in which some transcriptional factors regulate most of the other genes. Note here that genes regulated by the same transcriptional factor often show expression patterns similar to each other. Inference

methods generally perform poorly in discriminating genes of this type. One solution for this problem is to use some clustering technique to identify genes with similar expression patterns, group them together, and then infer the regulations between the

TABLE 1 | The performance of the proposed approach and the BS-LPM inference method evaluated with respect to the area under the recall-precision curve (AURPC).

| | Ecoli1 AVG ± STD | Ecoli2 AVG ± STD | Yeast1 AVG ± STD | Yeast2 AVG ± STD | Yeast3 AVG ± STD |
|-------------------------|---------------------|---------------------|---------------------|---------------------|---------------------|
| Proposed approach | 0.9042 ± 0.0402 | 0.9233 ± 0.0252 | 0.7140 ± 0.0515 | 0.4330 ± 0.0251 | 0.3563 ± 0.0386 |
| BS-LPM inference method | 0.8792 ± 0.0528 | 0.9154 ± 0.0207 | 0.6880 ± 0.0491 | 0.4151 ± 0.0207 | 0.3504 ± 0.0368 |

AVG and STD represent the averaged value of the AURPCs and its standard deviation, respectively.

TABLE 2 | The AURPCs of the proposed approach with different values for the parameter η .

| Parameter η | Ecoli1 AVG ± STD | Ecoli2 AVG ± STD | Yeast1 AVG ± STD | Yeast2 AVG ± STD | Yeast3 AVG ± STD |
|---------------------|---------------------|---------------------|---------------------|---------------------|---------------------|
| 1.000 | 0.8792 ± 0.0528 | 0.9154 ± 0.0207 | 0.6880 ± 0.0491 | 0.4151 ± 0.0207 | 0.3504 ± 0.0368 |
| 0.999 | 0.9043 ± 0.0402 | 0.9233 ± 0.0252 | 0.7140 ± 0.0515 | 0.4330 ± 0.0251 | 0.3563 ± 0.0386 |
| 0.995 | 0.9043 ± 0.0402 | 0.9233 ± 0.0252 | 0.7140 ± 0.0515 | 0.4330 ± 0.0251 | 0.3563 ± 0.0386 |
| 0.990 | 0.9042 ± 0.0402 | 0.9233 ± 0.0252 | 0.7140 ± 0.0515 | 0.4330 ± 0.0251 | 0.3563 ± 0.0386 |
| 0.950 | 0.8969 ± 0.0415 | 0.9159 ± 0.0272 | 0.6962 ± 0.0486 | 0.4269 ± 0.0249 | 0.3487 ± 0.0390 |
| 0.900 | 0.8898 ± 0.0430 | 0.9110 ± 0.0259 | 0.6832 ± 0.0459 | 0.4180 ± 0.0243 | 0.3412 ± 0.0390 |

Note that the proposed approach with $\eta = 1.000$ is equivalent to the BS-LPM inference method and the parameter $\eta = 1 - \frac{1}{N_g} = 0.990$ is our recommended setting.

clusters (see e.g., Kimura et al., 2005). There would thus be no need, in practical application, to detect hierarchical structures in networks with topological properties similar to Ecoli2.

As described in the section Assignment of Confidence Values to Regulations, this study uses the parameter η to combine the results from the BS-LPM inference method and those from the hierarchical random graph model. Therefore, we then checked the effect of the parameter η on the performance of the proposed approach. **Table 2** shows the AURPCs of our approach with different values for η . The experimental results indicate that, although the use of the hierarchical structure has an ability to improve the confidence values of regulations, we should not rely too much on it.

The performance of the proposed approach might depend on the number of the networks inferred by the BS-LPM inference method, N_g . Therefore, we also checked our setting of the parameter η , i.e., $\eta = 1 - \frac{1}{N_g}$, on the experiments with different numbers of N_g . **Figure 5** shows the AURPCs of the proposed approach with $N_g = 20, 50, 100$, and 200 on the problems of Yeast1. The figure indicates the reasonableness of our parameter setting.

As mentioned previously, our approach improves the confidence values of regulations by combining the probabilities evaluated by the BS-LPM inference method with those evaluated by the hierarchical random graph model. Note that this study obtains the hierarchical random graph model using the genetic networks inferred by the BS-LPM inference method. Therefore, the reasonableness of the extracted hierarchical structure depends on the accuracy of the inferred genetic networks. We investigated how the accuracy of the inferred genetic networks affected the performance of the proposed approach by performing experimental runs with variable amounts of time-series data applied to the problems of Yeast1. **Figure 6** plots

the averaged AURPC against the amount of time-series data. The plot shows that the use of the hierarchical structure has no negative effect on the inference ability, on average, even when the inferred networks are inaccurate.

4.2. Analysis of an Actual Network

We next applied the proposed approach to an experiment using actual data.

4.2.1. Experimental Setup

This experiment analyzed an ErbB-receptor-mediated regulatory network of transcription factors in normal human epidermal keratinocytes. The network consisted of 29 components, i.e., three receptors (EGFR, ErbB2, and ErbB3), seven signal transducer proteins (ERK, PI3K, AKT, STAT3, PLCg, PKCd, and c-SRC), the phosphorylated forms of the three receptors and the seven signal transducer proteins, and seven transcription factors (c-FOS, FRA1, FRA2, JUNB, c-JUN, JUND, and c-MYC). Time-series data consisting of 14 measurements of the 29 components were measured by Saeki et al. (2012). Lacking sufficient data, we inferred the target network using the following a priori knowledge: (i) none of the receptors or signaling proteins are affected by other receptors or signaling proteins; (ii) none of the transcription factors are affected by receptors, signaling proteins, or phosphorylated forms of receptors; (iii) none of the phosphorylated receptors or phosphorylated signaling proteins are affected by other receptors, signaling proteins, or transcription factors; (iv) every component of this system regulates itself; (v) every protein regulates its own phosphorylated form. We employed this knowledge according to the biological knowledge that phosphorylated forms of signaling proteins and receptors can form cascades to transduce extracellular signals to transcription factors (Alberts et al., 2008). Based

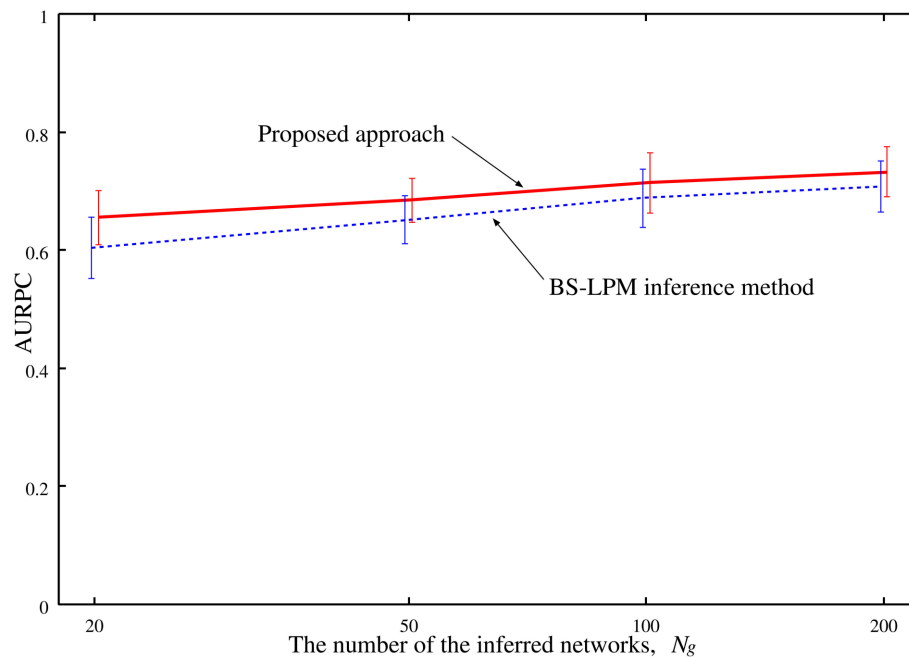


FIGURE 5 | The averaged AURPCs of the proposed approach (solid line) and the BS-LPM inference method (dotted line) plotted against the number of the inferred networks, N_g .

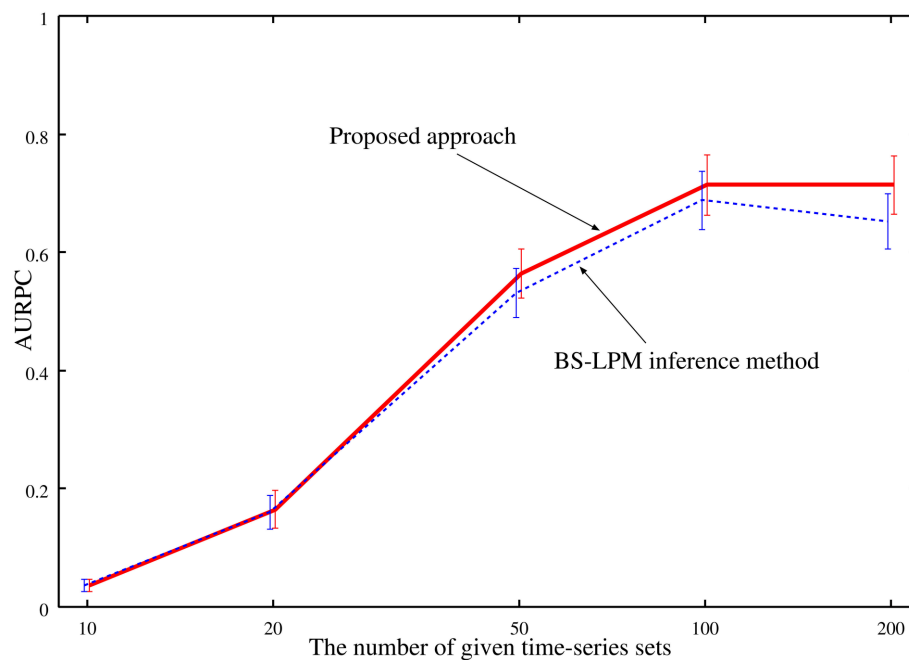


FIGURE 6 | The averaged AURPCs of the proposed approach (solid line) and the BS-LPM inference method (dotted line) in the experiments where different numbers of time-series sets were given.

on the knowledge (i), for example, we prohibited inferring the regulation of EGFR from ErbB2. We used the technique proposed by Kimura et al. (2009b) in order to introduce the knowledge described above into the inference method. By

introducing this a priori knowledge, we reduced the degree-of-freedom of the network model. The other experimental conditions were the same as those in the section Analysis of Dream3 Networks.

4.2.2. Results

The network of the regulations with confidence values exceeding 0.25 is shown in **Figure 8A**. The network obtained contained 135 regulations, but 17 were regulations of the proteins from their phosphorylated forms or vice versa, which probably made them trivial. We still lack a detailed understanding of the regulatory network used for this study, which consisted of proteins and their phosphorylated forms. We therefore compared the inferred network with a protein network consisting of the proteins alone (**Figure 7**). We obtained this protein network from the STRING database (<http://string-db.org/>) (Szklarczyk et al., 2015). The comparison results indicate that 77 of the 135 inferred regulations were reasonable, since the interactions between the corresponding proteins have been reportedly confirmed.

The proposed approach extracted a hierarchical structure of the target network from the networks inferred by the BS-LPM inference method. The extracted hierarchical structure is shown in **Figure 8B**. As the figure indicates, the network contained three clusters, i.e., clusters 1, 2, and 3. Clusters 1 and 2 contained the transcription factors, the downstream components of the target pathway. Cluster 3 mainly contained the upstream components. The phosphorylated ERK and the phosphorylated STAT3's, none of which belonged to any cluster, were intermediate components thought to regulate the transcription factors (Saeki et al., 2012). Although imperfect, the hierarchical structure

obtained seemed to reflect the actual structure of the target pathway. We thus think that the hierarchical random graph model obtained can be used to assess the reliability of the inferred network and/or to understand the structure of the target network.

As mentioned before, our approach is highly dependent on the BS-LPM inference method. The inferred network was therefore almost the same as that obtained only from the BS-LPM inference method. Our approach improves the confidence values of the regulations using the hierarchical random graph model obtained. We know, for example, that the phosphorylated ERK and phosphorylated STAT3 regulate each other (Gao and Horvath, 2008). In this experiment, these regulations were inferred by both the proposed approach and the BS-LPM inference method. The BS-LPM inference method assigned a confidence value of 0.29 to the regulation of the phosphorylated STAT3 by the phosphorylated ERK, and assigned the same value to four other regulations. Our approach, on the other hand, assigned a confidence value of 0.2922 to the regulation of the phosphorylated STAT3 by the phosphorylated ERK, a value superior the confidence values assigned to the same four other regulations. This feature of our approach could be useful for reducing the efforts of biologists to experimentally validate inferred regulations.

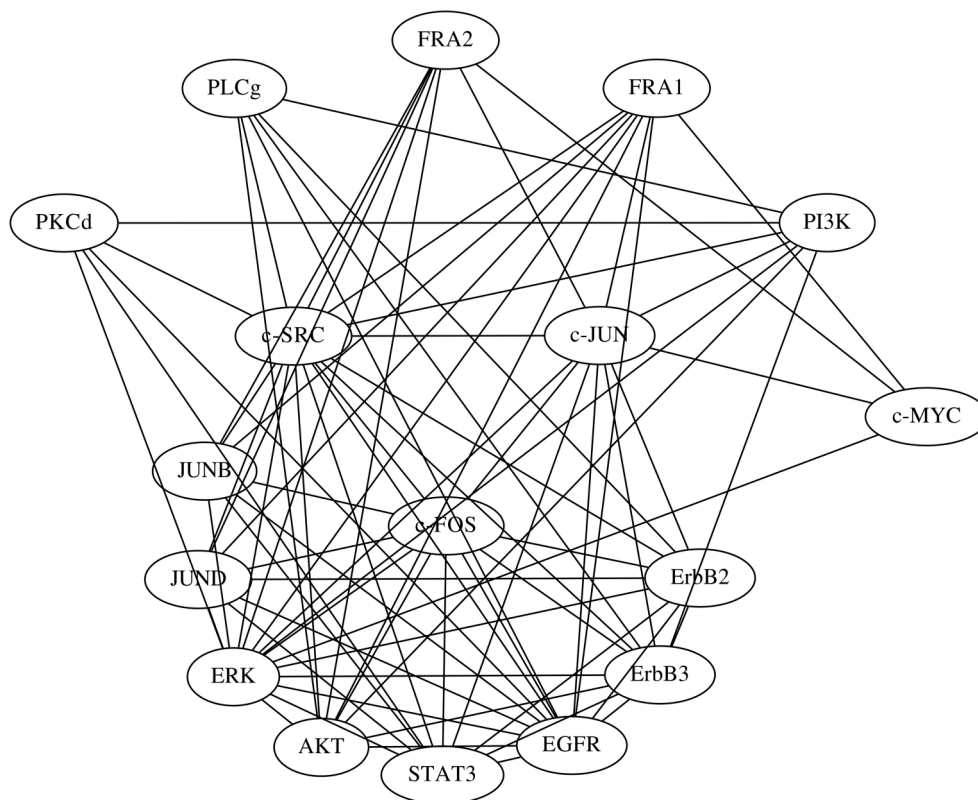


FIGURE 7 | The protein network obtained from the STRING database. Edges represent protein-protein interactions that have been reportedly confirmed by biochemical experiments.

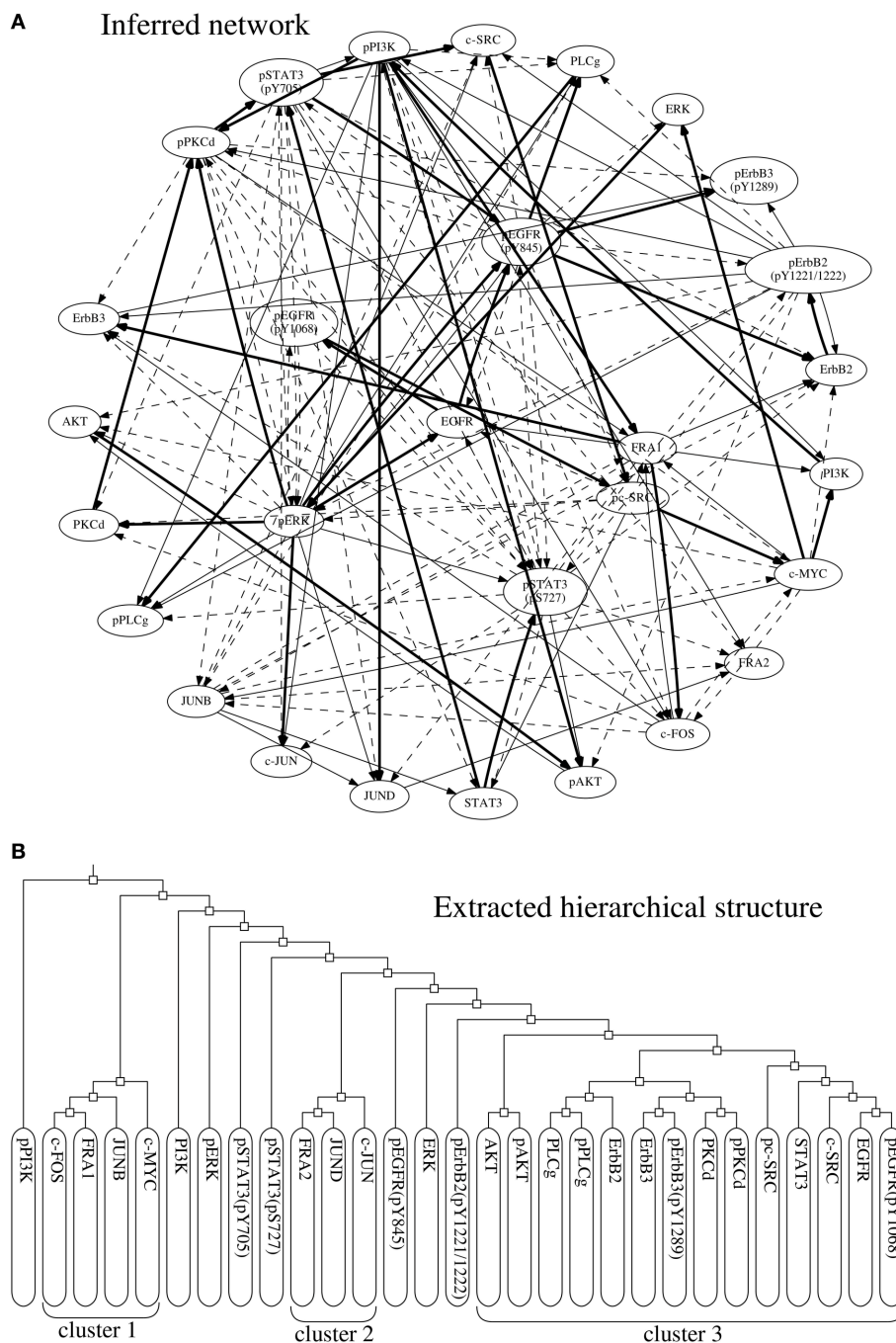


FIGURE 8 | (A) The network of regulations with confidence values exceeding 0.25. Bold, solid, and dotted lines represent regulations with confidence values exceeding 0.75, 0.5, and 0.25, respectively. **(B)** The hierarchical structure extracted from the inferred networks.

5. CONCLUSION

In this paper, we have proposed an approach for inferring a more reasonable genetic network by utilizing the hierarchical structures in genetic networks. The first step of this new approach is to infer multiple genetic networks from the given gene expression data. In this study, we took this step using the BS-LPM

inference method (Kimura et al., 2010). The next steps in our approach are to extract the hierarchical structure in the target network from the genetic networks generated in the first step, and then to use the extracted hierarchical structure to compute the confidence values of the regulations. Our experimental results showed that the use of the hierarchical structure improves the confidence values of the regulations. As mentioned in the section

Assignment of Confidence Values to Regulations, however, this study used the obtained hierarchical structure to rank the regulations that are assigned the same probability by the BS-LPM inference method. When there are no regulations that have the same bootstrap probability, therefore, the use of the hierarchical structure does not work. In our future work, thus, we must improve this drawback.

The approach proposed in this study consists of a BS-LPM inference method and a method for detecting hierarchical structures. The BS-LPM inference method is a combination of the LPM-based inference method (Kimura et al., 2009a) and the bootstrap method. We have the freedom, however, to use any inference method in place of the LPM-based inference method. Meanwhile, several investigators have proposed other inference methods that are capable of assigning confidence values to regulations without the use of the bootstrap method (see e.g., Huynh-Thu et al., 2010). The use of the hierarchical structure may also be effective in improving the performance of these methods.

Several inference methods that utilize a priori knowledge about the properties of genetic networks have been already proposed (see e.g., Kikuchi et al., 2003; Daisuke and Horton,

2006). These methods use the a priori knowledge during the genetic network inference. We could say, on the other hand, that the proposed approach uses the a priori knowledge after inferring genetic networks. Our experimental results proved that, even after the genetic network inference, the use of the a priori knowledge has an ability to improve the confidence values of regulations. Thus, although the improvement done by the proposed approach was very small, our framework might enable us to use other types of a priori knowledge that are currently difficult to utilize.

AUTHOR CONTRIBUTIONS

SK designed the method and performed the experiments. MT implemented some parts of the proposed algorithm. MOH supervised the biological aspect of this work. All authors read and approved the manuscript.

ACKNOWLEDGMENTS

This work has been supported by JSPS KAKENHI Grant Number 26330275.

REFERENCES

- Alberts, B., Johnson, A., Lewis, J., Raff, M., Roberts, K., and Walter, P. (2008). *Molecular Biology of the Cell, 5th Edn.* New York, NY: Garland Science.
- Chib, S., and Greenberg, E. (1995). Understanding the Metropolis-Hastings algorithm. *Am. Statist.* 49, 327–335.
- Chou, I. C., and Voit, E. O. (2009). Recent developments in parameter estimation and structure identification of biochemical and genomic systems. *Math. Biosci.* 219, 57–83. doi: 10.1016/j.mbs.2009.03.002
- Clauset, A., Moore, C., and Newman, M. E. J. (2008). Hierarchical structure and the prediction of missing links in networks. *Nature* 453, 98–101. doi: 10.1038/nature06830
- Daisuke, T., and Horton, P. (2006). Inference of scale-free networks from gene expression time series. *J. Bioinform. Comput. Biol.* 4, 503–514. doi: 10.1142/S0219720006001886
- Efron, B. (1979). Bootstrap methods: another look at the Jackknife. *Ann. Statist.* 7, 1–26.
- Gao, Q., and Horvath, T. L. (2008). Cross-talk between estrogen and leptin signaling in the hypothalamus. *Am. J. Physiol.* 294, E817–E826. doi: 10.1152/ajpendo.00733.2007
- Hecker, M., Lambeck, S., Toepfer, S., van Someren, E., and Guthke, R. (2009). Gene regulatory network inference: data integration in dynamic models – a review. *Biosystems* 96, 86–103. doi: 10.1016/j.biosystems.2008.12.004
- Huynh-Thu, V. A., Irrthum, A., Wehenkel, L., and Geurts, P. (2010). Inferring regulatory networks from expression data using tree-based methods. *PLoS ONE* 5:e12776. doi: 10.1371/journal.pone.0012776
- Kikuchi, S., Tominaga, D., Arita, M., Takahashi, K., and Tomita, M. (2003). Dynamic modeling of genetic networks using genetic algorithm and S-system. *Bioinformatics* 19, 643–650. doi: 10.1093/bioinformatics/btg027
- Kimura, S., Ide, K., Kashiwara, A., Kano, M., Hatakeyama, M., Masui, R., et al. (2005). Inference of S-system models of genetic networks using a cooperative coevolutionary algorithm. *Bioinformatics* 21, 1154–1163. doi: 10.1093/bioinformatics/bti071
- Kimura, S., Nakayama, S., and Hatakeyama, M. (2009a). Genetic network inference as a series of discrimination tasks. *Bioinformatics* 25, 918–925. doi: 10.1093/bioinformatics/btp072
- Kimura, S., Shiraishi, Y., and Hatakeyama, M. (2009b). Inference of genetic networks using linear programming machines: application of a priori knowledge. *Proc. Int. Joint Conf. Neural Netw.* 1617–1624. doi: 10.1109/ijcnn.2009.5178679
- Kimura, S., Shiraishi, Y., and Okada, M. (2010). Inference of genetic networks using LPMs: assessment of confidence values of regulations. *J. Bioinform. Comput. Biol.* 8, 661–677. doi: 10.1142/S0219720010004859
- Kimura, S., and Okada-Hatakeyama, M. (2015). Improvement of reliabilities of regulations using a hierarchical structure in a genetic network. *Proc. Int. Joint Conf. Neural Netw.* 485–491. doi: 10.1126/science.220.4598.671
- Kirkpatrick, S., Gelatt C. D. Jr., and Vecchi, M. P. (1983). Optimization by simulated annealing. *Science* 220, 671–680.
- Larrañaga, R., Calvo, B., Santana, R., Bielza, C., Galdiano, J., Inza, I., et al. (2006). Machine learning in bioinformatics. *Brief. Bioinform.* 7, 86–112. doi: 10.1093/bib/bbk007
- Lim, N., Senbabaoğlu, Y., Michailidis, G., and d'Alché-Buc, F. (2013). OKVAR-boost: a novel boosting algorithm to infer nonlinear dynamics and interactions in gene regulatory networks. *Bioinformatics* 29, 1416–1423. doi: 10.1093/bioinformatics/btt167
- Prill, R. J., Marbach, D., Saez-Rodriguez, J., Sorger, P. K., Alexopoulos, L. G., Xue, X., et al. (2010). Towards a rigorous assessment of systems biology models: the DREAM3 challenges. *PLoS ONE* 5:e9202. doi: 10.1371/journal.pone.0009202
- Saeki, Y., Nagashima, T., Kimura, S., and Okada-Hatakeyama, M. (2012). An ErbB receptor-mediated AP-1 regulatory network is modulated by STAT3 and c-MYC during calcium-dependent keratinocyte differentiation. *Exp. Dermatol.* 21, 293–298. doi: 10.1111/j.1600-0625.2012.01453.x
- Szklarczyk, D., Franceschini, A., Wyder, S., Forslund, K., Heller, D., Huerta-Cepas, J., et al. (2015). STRING v10: protein-protein interaction networks, integrated over the tree of life. *Nucleic Acids Res.* 43, D447–D452. doi: 10.1093/nar/gku1003
- Voit, E. O. (2000). *Computational Analysis of Biochemical Systems*. Cambridge: Cambridge University Press.

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2016 Kimura, Tokuhisa and Okada-Hatakeyama. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



Book Review: Python Programming for Biology

Alberto Marin-Sanguino *

Specialty Division for Systems Biotechnology, Faculty for Mechanical Engineering, Technical University of Munich, Garching, Germany

Keywords: Python, systems biology, computational biology, bioinformatics, programming languages

**A book review on
Python Programming for Biology: Bioinformatics and Beyond**

Edited by T. J. Stevens and W. Boucher, Cambridge University Press, 2015. ISBN: 978-0-521-72009-0

Python is an excellent language for scientists and can be used at many levels: Simple scripts for quick and dirty calculations, big programs implementing complex data models, taking advantage of its powerful libraries for number crunching or simply as “glue” to bind together more specialized modules written in C or Fortran. Such a versatility may be an obstacle to find an entry point for the newcomer. Stevens and Boucher (2015) take computational Biology in the widest sense of the term to write an extensive introduction to Python and provide an overview of its main scientific libraries.

The reader will find a gentle but terse introduction covering the basics in seventy pages before moving on to practical topics like reading relevant file formats (FASTA, pdb, ...) From that point on, programming concepts like data models or object orientation are introduced through practical examples. The rest of the book is composed by an assortment of chapters that cover a wide diversity of topics and can be read independently or used as a reference. Some of these chapters are focused on specific biological topics such as macromolecular structures, sequence alignments, array data and high-throughput sequencing. The relevant libraries are introduced along one or more of these chapters to get the user up to speed. Other chapters are centered on a technique or discipline such as image processing, machine learning, probability or statistics. In all cases, a brief introduction on the topic is followed by a series of examples on how to get started. The selection of topics offers a good compromise, presenting the general principles as well as useful recipes without overwhelming the reader with excessive detail. Getting the reader up to speed in a very short time seems to be the key premise. Finally, advanced topics like parallelization and interfacing with C will point the reader to the next level. The book also provides a good overview of the main libraries with immediate applications to biology, although some readers may miss a chapter on pandas.

This book is a good choice for researchers who want to migrate to Python or Ph.D. students about to get started a computational biology or bioinformatics project. Biologists without programming experience may prefer to start with a more gentle and maybe shorter introduction, but those with previous experience with software packages like MATLAB or R will also find this book to be a fast lane to Python. Clear explanations of the biological background will also make this book accessible to scientists intending to move into biology from other disciplines.

AUTHOR CONTRIBUTIONS

The author confirms being the sole contributor of this work and approved it for publication.

OPEN ACCESS

Edited and reviewed by:

Pierre De Meyts,
De Meyts R&D Consulting, Belgium

***Correspondence:**

Alberto Marin-Sanguino
a.marin@lrz.tu-muenchen.de

Specialty section:

This article was submitted to
Systems Biology,
a section of the journal
Frontiers in Genetics

Received: 23 March 2016

Accepted: 08 April 2016

Published: 26 April 2016

Citation:

Marin-Sanguino A (2016) Book
Review: Python Programming for
Biology. *Front. Genet.* 7:66.
doi: 10.3389/fgene.2016.00066

FUNDING

AMS acknowledges funding from the German Federal Ministry of Education and Research (BMBF), e:Bio initiative 0316197.

REFERENCES

Stevens, T. J., and Boucher, W. (2015). *Python Programming for Biology: Bioinformatics and Beyond*. Cambridge University Press.

Conflict of Interest Statement: The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2016 Marin-Sanguino. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



Book Review: Computing for Biologists: Python Programming and Principles

Alberto Marin-Sanguino *

Specialty Division for Systems Biotechnology, Faculty for Mechanical Engineering, Technische Universität München, München, Germany

Keywords: python, computational biology, systems biology, programming languages, computing

A book review on

Computing for Biologists: Python Programming and Principles

Edited by Ran Libeskind-Hadas and Eliot Bush, Cambridge, New York: Cambridge University Press, 2014. ISBN: 978-1-107-64218-8

Fifteen years ago a student could go through the whole curriculum in any of the life sciences without ever switching a computer on but, as bioinformatics and systems biology gained weight in the field, biologists became first users and then developers of increasingly sophisticated computer programs. Even for professionals far away from these disciplines, it is nowadays unthinkable to plan an experiment without first checking some databases, designing some primers or analyzing a sequence. These changes have created a need to include basic courses in the biology curriculum where students with little or no inclination toward computer science have to learn the ropes of programming. The main challenge when teaching such a course is to find a good balance between presenting examples that are complex enough to motivate the students and simple enough to be accessible for them. This book is precisely a guide for such an introductory course. Using python as a first programming language and assuming no previous knowledge, the authors follow a practical approach to teach the basic programming skills.

The book is structured in four parts, preceded by a twenty pages tutorial on the basics of python. The examples in each of the first three parts revolve around a unifying biological theme, which turns each of them into a simple yet interesting project. Abstract computational concepts like recursion or memoization are introduced as they are needed to solve diverse problems. At the end of each part, a problem is formulated to solve a real case study using the material covered so far. Abundant examples, additional explanations on these problems and source code with the solutions are available through the companion website. The first part starts with simple tasks like computing GC content of a DNA sequence or converting DNA to its corresponding mRNA. Through these simple examples, the usage of different data types is introduced as well as the basics of flow control and functions. The rest of the first part is dedicated to the consolidate knowledge on these basic concepts understanding the general organization of a program. These concepts are then used to find ORFs in a genome and sequences that are associated with pathogenicity in *Salmonella*. The second part covers sequence alignments, building up toward finding homology between genes and then chromosomes. It is all exemplified by comparing X chromosome in humans and Z chromosome in chicken. The third part covers phylogenetic trees and ends with mitochondrial DNA comparison between humans and neanderthals. Finally, a more heterogeneous part closes the book presenting three different examples that depart from the rest: RNA folding, finding gene regulation networks and genetic algorithms.

OPEN ACCESS

Edited by:

Firas H. Kobeissy,
University of Florida, USA

Reviewed by:

Pierre Khoeiry,
American University of Beirut,
Lebanon

*Correspondence:

Alberto Marin-Sanguino
A.Marin@lrz.tu-muenchen.de

Specialty section:

This article was submitted to
Systems Biology,
a section of the journal
Frontiers in Genetics

Received: 23 March 2016

Accepted: 25 April 2016

Published: 11 May 2016

Citation:

Marin-Sanguino A (2016) Book
Review: Computing for Biologists:
Python Programming and Principles.
Front. Genet. 7:86.
doi: 10.3389/fgene.2016.00086

The text is extremely well written, with clear explanations and interesting examples. The pace is slow but entertaining, ensuring that the student can keep up with the step by step explanations while sitting in front of the computer and trying things right away. No previous knowledge is assumed, and the most basic programming concepts are explained from scratch, stopping to indicate possible pitfalls and preparing the reader for potential difficulties. There are many introductory texts to Python, including some aimed to biologists. What makes this book different is that it does not focus on teaching a particular programming language or some useful algorithms. The authors present biological problems and keep the attention focused on them. Python is just taken as a vehicle to introduce very abstract concepts by example. This makes the book valuable as a guide for an undergraduate course. Even if the students never use python after the course is over, they will have acquired the basic programming skills they may need. I haven't found a book that follows this approach so successfully since "Beginning Perl for Bioinformatics" by James Tisdall, 15 years ago.

AUTHOR CONTRIBUTIONS

The author confirms being the sole contributor of this work and approved it for publication.

FUNDING

This work was funded by the German Ministry of Education and Research (BMBF) through the e:Bio initiative (project OpHeLia—0316197).

Conflict of Interest Statement: The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2016 Marin-Sanguino. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

Advantages of publishing in Frontiers



OPEN ACCESS

Articles are free to read
for greatest visibility
and readership



FAST PUBLICATION

Around 90 days
from submission
to decision



HIGH QUALITY PEER-REVIEW

Rigorous, collaborative,
and constructive
peer-review



TRANSPARENT PEER-REVIEW

Editors and reviewers
acknowledged by name
on published articles

Frontiers

Avenue du Tribunal-Fédéral 34
1005 Lausanne | Switzerland

Visit us: www.frontiersin.org

Contact us: info@frontiersin.org | +41 21 510 17 00



REPRODUCIBILITY OF RESEARCH

Support open data
and methods to enhance
research reproducibility



DIGITAL PUBLISHING

Articles designed
for optimal readership
across devices



FOLLOW US

@frontiersin



IMPACT METRICS

Advanced article metrics
track visibility across
digital media



EXTENSIVE PROMOTION

Marketing
and promotion
of impactful research



LOOP RESEARCH NETWORK

Our network
increases your
article's readership