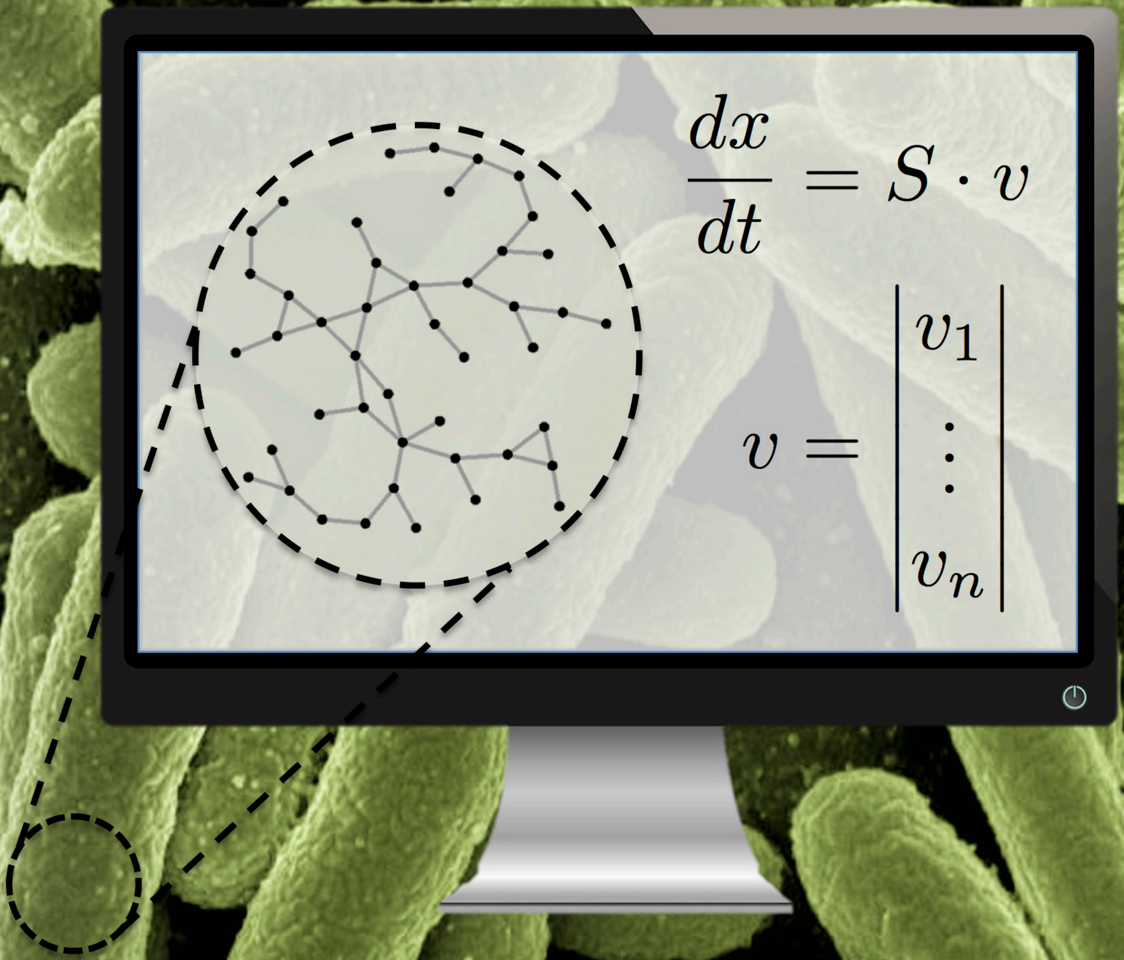
CURRENT CHALLENGES IN MODELING CELLULAR METABOLISM

EDITED BY: Daniel Machado, Kai H. Zhuang, Nikolaus Sonnenschein and
Markus J. Herrgård

PUBLISHED IN: Frontiers in Bioengineering and Biotechnology and
Frontiers in Cell and Developmental Biology


$\frac{dx}{dt} = S \cdot v$
$v = \begin{vmatrix} v_1 \\ \vdots \\ v_n \end{vmatrix}$

frontiers Research Topics

ISSN 1664-8714
ISBN 978-2-88919-754-5
DOI 10.3389/978-2-88919-754-5

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.

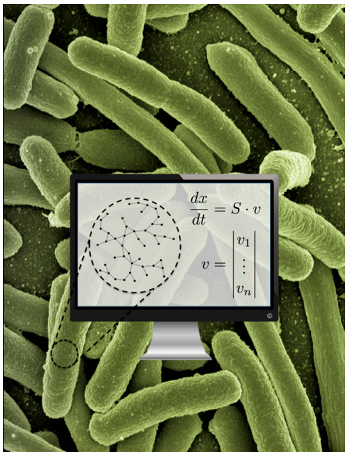
By applying the most advanced information technologies, Frontiers is catapulting scholarly publishing into a new generation.

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**

CURRENT CHALLENGES IN MODELING CELLULAR METABOLISM

Topic Editors:
Daniel Machado, University of Minho, Portugal
Kai H. Zhuang, Technical University of Denmark, Denmark
Nikolaus Sonnenschein, Technical University of Denmark, Denmark
Markus J. Herrgård, Technical University of Denmark, Denmark


$\frac{dx}{dt} = S \cdot v$
$v = \begin{vmatrix} v_1 \\ \vdots \\ v_n \end{vmatrix}$

Mathematical modeling and computational methods provide an essential framework to unravel the complexity of cellular metabolism.

Image adapted from: https://pixabay.com/en/koli-bacteria-escherichia-coli-123081/ and https://pixabay.com/en/monitor-isolated-display-white-313011/

Mathematical and computational models play an essential role in understanding the cellular metabolism. They are used as platforms to integrate current knowledge on a biological system and to systematically test and predict the effect of manipulations to such systems. The recent advances in genome sequencing techniques have facilitated the reconstruction of genome-scale metabolic networks for a wide variety of organisms from microbes to human cells. These models have been successfully used in multiple biotechnological applications.

Despite these advancements, modeling cellular metabolism still presents many challenges. The aim of this Research Topic is not only to expose and consolidate the state-of-the-art in metabolic modeling approaches, but also to push this frontier beyond the current edge through the introduction of innovative solutions.

The articles presented in this e-book address some of the main challenges in the field, including the integration of different modeling formalisms, the integration of heterogeneous data sources into metabolic models, explicit representation of other biological processes during phenotype simulation, and standardization efforts in the representation of metabolic models and simulation results.

Citation: Machado, D., Zhuang, K. H., Sonnenschein, N., Herrgård, M. J., eds. (2016). Current Challenges in Modeling Cellular Metabolism. Lausanne: Frontiers Media. doi: 10.3389/978-2-88919-754-5

Table of Contents

frontiers
in Bioengineering and Biotechnology

EDITORIAL
published: 26 November 2015
doi: 10.3389/fbioe.2015.00193

Editorial: Current Challenges in Modeling Cellular Metabolism

Daniel Machado[1], Kai H. Zhuang[2], Nikolaus Sonnenschein[2] and Markus J. Herrgård[2]*

[1] Centre of Biological Engineering, University of Minho, Braga, Portugal, [2] The Novo Nordisk Foundation Center for Biosustainability, Technical University of Denmark, Hørsholm, Denmark

Keywords: metabolism, modeling formalisms, metabolic networks, genome-scale modeling, kinetic modeling

OPEN ACCESS

Edited and reviewed by:
*Raina Robeva,
Randolph-Macon College, USA;
Sweet Briar College, USA*

****Correspondence:***
*Daniel Machado
dmachado@deb.uminho.pt*

Specialty section:
This article was submitted to Systems Biology, a section of the journal Frontiers in Bioengineering and Biotechnology

Received: *29 October 2015*
Accepted: *09 November 2015*
Published: *26 November 2015*

Citation:
Machado D, Zhuang KH, Sonnenschein N and Herrgård MJ (2015) Editorial: Current Challenges in Modeling Cellular Metabolism. Front. Bioeng. Biotechnol. 3:193. doi: 10.3389/fbioe.2015.00193

Metabolism is a core process of every cell providing the energy and building blocks for all other biological processes. Mathematical models and computational tools have become essential for unraveling the complexity of cellular metabolism (Heinemann and Sauer, 2010). Models integrate current knowledge on a biological system in an unambiguous manner and allow simulating cellular responses to genetic and environmental perturbations. Advances in genome sequencing and annotation have facilitated the reconstruction of genome-scale metabolic models for hundreds of organisms, which are currently used in various applications ranging from human health to industrial biotechnology (Bordbar et al., 2014).

Despite these advancements, there are still major challenges in modeling cellular metabolism at the genome scale. These include the reconciliation of different modeling approaches, the integration of metabolic models with models of other biological processes, the interpretation of heterogeneous data sources using models, and the adoption of suitable standards for model sharing. The aim of this Research Topic is to present state-of-the-art methods that aim to overcome these challenges and push this frontier to a new edge.

Starting from the most fundamental aspect of biochemical reactions, Cannon (2014) reviews the historical perspective of thermodynamics as a major driving force in the evolution of life and presents a primer on statistical thermodynamics. The author then provides examples of thermodynamic analysis of small metabolic pathways, highlighting future directions for integration of thermodynamics and large-scale modeling.

The most common approach to build a metabolic model is bottom-up reconstruction, where individual reactions for a given organism are identified (through genome annotation and literature data) and retrieved from biochemical databases. This approach is mostly limited by the current knowledge on enzymes with annotated functions. The alternative (termed top-down) approach is to infer the underlying network structure by reverse engineering of metabolome data. Çakir and Khatibipour (2014) compare these two approaches, reviewing available methods for both cases and providing pointers toward the reconciliation of these strategies.

Once a model is built, it can be used to simulate the metabolic phenotype under different conditions and subsequently compared with *in vivo* results for validation and refinement. Phenotype microarrays currently allow high-throughput assessment of metabolic responses to multiple experimental conditions. Chaiboonchoe et al. (2014) present an optimization of the Biolog phenotyping protocol for metabolic profiling of microalgae. The experimental results are used to expand and refine a genome-scale model of the alga *Chlamydomonas reinhardtii* to include the utilization of carbon and nitrogen sources not present in the original model.

Choosing a modeling formalism requires a compromise between model size and detail (Machado et al., 2011). Constraint-based models have gained popularity for their scalability to the genome scale. However, when insight of intracellular dynamics is required, kinetic models become the obvious choice. Petri nets, with their varied extensions, offer an intermediate level of compromise, allowing structural network analysis and, to some extent, dynamic analysis. Hartmann and Schreiber (2015) present a unified graph formalism and implement transformation operations to convert from the unified model to any specific formalism. The authors provide an example of integrated analysis using different formalisms in a unified model of sucrose breakdown in the potato tuber.

Current *omics* technologies allow unprecedented quantification of different types of cellular components including RNA transcript, protein, and metabolite levels. Machado et al. (2015) use a multi-*omics* dataset of *Escherichia coli* to analyze the contribution of allosteric regulation in controlling central carbon metabolism. Given the role of this type of control in response to different perturbations, the authors present a new simulation method to account for allosteric interactions in the determination of steady-state flux distributions. This is the first constraint-based method to account for allosteric regulation.

Next-generation sequencing is another example of technology pushing the limits of biological discovery. Understanding how genetic variants affect metabolic phenotype is fundamental in diverse areas, such as the study of disease mechanisms and the engineering of microbial cell factories. Cardoso et al. (2015) review available methods to predict the effect of genetic variations in protein function and expression. Integrating these methods with genome-scale metabolic modeling creates the potential for mechanistically predicting the consequences of genetic variation in the cellular phenotype, which is currently not possible with the statistical approaches used in genome-wide association studies.

Microbial strain design is a common application of genome-scale models as the combinatorial explosion of possible genetic manipulations demands efficient optimization methods. Stanford et al. (2015) address the problem of butanol production in *E. coli* using a new strain design method, RobOKoD, that combines gene over/underexpression with gene knockouts, showing good agreement with experimental data. Khodayari et al. (2015) analyze the case of succinate overproduction in *E. coli* using k-OptForce, the first strain design method that accounts for integrated simulation of kinetic and constraint-based models. This enables strain design at the genome scale while accounting for regulation mechanisms in central carbon pathways, such as feedback inhibition. The authors observe decreased prediction accuracy when the kinetic model is applied in experimental conditions that differ from those used for parameter estimation, highlighting the importance of reparameterization of kinetic models for the conditions used in the production setting.

Last but not least, modeling the complexity of cellular metabolism is an iterative refinement process that cannot be accomplished without a community effort. The ability to share models using suitable standards is of paramount importance (Ebrahim et al., 2015). Dräger and Palsson (2014) present a comprehensive review of standardization efforts in Systems Biology, including standards for model representation, model visualization, minimum information requirements, and suitable ontologies. This review also covers public model databases, conversion tools, simulation software, and standards for publication of simulation results. Adoption of these standards is essential to ensure reusability of models and reproducibility of results.

The work presented in this Research Topic addresses many of the current gaps in the field with innovative solutions. Closing these gaps provides a stepping stone for the challenges to come. The future of metabolic modeling already holds exciting opportunities with a new generation of models that include protein structures, gene expression pathways, and even whole-cell models (King et al., 2015).

AUTHOR CONTRIBUTIONS

All authors have read and revised the manuscript.

ACKNOWLEDGMENTS

The authors would like to thank the financial support from the Novo Nordisk Foundation.

REFERENCES

Bordbar, A., Monk, J. M., King, Z. A., and Palsson, B. O. (2014). Constraint-based models predict metabolic and associated cellular functions. *Nat. Rev. Genet.* 15, 107–120. doi:10.1038/nrg3643

Çakir, T., and Khatibipour, M. J. (2014). Metabolic network discovery by top-down and bottom-up approaches and paths for reconciliation. *Front. Bioeng. Biotechnol.* 2:62. doi:10.3389/fbioe.2014.00062

Cannon, W. R. (2014). Concepts, challenges and successes in modeling thermodynamics of metabolism. *Front. Bioeng. Biotechnol.* 2:53. doi:10.3389/fbioe.2014.00053

Cardoso, J. G., Andersen, M. R., Herrgård, M. J., and Sonnenschein, N. (2015). Analysis of genetic variation and potential applications in genome-scale metabolic modeling. *Front. Bioeng. Biotechnol.* 3:13. doi:10.3389/fbioe.2015.00013

Chaiboonchoe, A., Dohai, B. S., Cai, H., Nelson, D. R., Jijakli, K., and Salehi-Ashtiani, K. (2014). Microalgal metabolic network model refinement through high throughput functional metabolic profiling. *Front. Bioeng. Biotechnol.* 2:68. doi:10.3389/fbioe.2014.00068

Dräger, A., and Palsson, B. O. (2014). Improving collaboration by standardization efforts in systems biology. *Front. Bioeng. Biotechnol.* 2:61. doi:10.3389/fbioe.2014.00061

Ebrahim, A., Almaas, E., Bauer, E., Bordbar, A., Burgard, A. P., Chang, R. L., et al. (2015). Do genome-scale models need exact solvers or clearer standards? *Mol. Syst. Biol.* 11, 831. doi:10.15252/msb.20156157

Hartmann, A., and Schreiber, F. (2015). Integrative analysis of metabolic models – from structure to dynamics. *Front. Bioeng. Biotechnol.* 2:91. doi:10.3389/fbioe.2014.00091

Heinemann, M., and Sauer, U. (2010). Systems biology of microbial metabolism. *Curr. Opin. Microbiol.* 13, 337–343. doi:10.1016/j.mib.2010.02.005

Khodayari, A., Chowdhury, A., and Maranas, C. D. (2015). Succinate overproduction: a case study of computational strain design using a comprehensive *Escherichia coli* kinetic model. *Front. Bioeng. Biotechnol.* 2:76. doi:10.3389/fbioe.2014.00076

King, Z. A., Lloyd, C. J., Feist, A. M., and Palsson, B. O. (2015). Next-generation genome-scale models for metabolic engineering. *Curr. Opin. Biotechnol.* 35, 23–29. doi:10.1016/j.copbio.2014.12.016

Machado, D., Costa, R. S., Rocha, M., Ferreira, E. C., Tidor, B., and Rocha, I. (2011). Modeling formalisms in systems biology. *AMB Express* 1, 1–14. doi:10.1186/2191-0855-1-45

Machado, D., Herrgård, M. J., and Rocha, I. (2015). Modeling the contribution of allosteric regulation for flux control in the central carbon metabolism of *E. coli*. *Front. Bioeng. Biotechnol.* 3:154. doi:10.3389/fbioe.2015.00154

Stanford, N. J., Swainston, N., and Millard, P. (2015). RobOKoD: microbial strain design for (over)production of target compounds. *Front. Cell Dev. Biol.* 3:17. doi:10.3389/fcell.2015.00017

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.

frontiers in
BIOENGINEERING AND BIOTECHNOLOGY

TECHNOLOGY REPORT
published: 26 November 2014
doi: 10.3389/fbioe.2014.00053

Concepts, challenges, and successes in modeling thermodynamics of metabolism

William R. Cannon*

Computational Biology and Bioinformatics Group, Biological Sciences Division, Pacific Northwest National Laboratory, Richland, WA, USA

Edited by:
Daniel Machado, University of Minho, Portugal

Reviewed by:
Keng Cher Soh, Ciris Energy, USA
Daniel Craig Zielinski, University of California San Diego, USA

****Correspondence:***
William R. Cannon, Computational Biology and Bioinformatics Group, Biological Sciences Division, Pacific Northwest National Laboratory, 902 Battelle Blvd, Richland, WA 99352, USA
e-mail: william.cannon@pnnl.gov

The modeling of the chemical reactions involved in metabolism is a daunting task. Ideally, the modeling of metabolism would use kinetic simulations, but these simulations require knowledge of the thousands of rate constants involved in the reactions. The measurement of rate constants is very labor intensive, and hence rate constants for most enzymatic reactions are not available. Consequently, constraint-based flux modeling has been the method of choice because it does not require the use of the rate constants of the law of mass action. However, this convenience also limits the predictive power of constraint-based approaches in that the law of mass action is used only as a constraint, making it difficult to predict metabolite levels or energy requirements of pathways. An alternative to both of these approaches is to model metabolism using simulations of states rather than simulations of reactions, in which the state is defined as the set of all metabolite counts or concentrations. While kinetic simulations model reactions based on the likelihood of the reaction derived from the law of mass action, states are modeled based on likelihood ratios of mass action. Both approaches provide information on the energy requirements of metabolic reactions and pathways. However, modeling states rather than reactions has the advantage that the parameters needed to model states (chemical potentials) are much easier to determine than the parameters needed to model reactions (rate constants). Herein, we discuss recent results, assumptions, and issues in using simulations of state to model metabolism.

Keywords: statistical thermodynamics, metabolism, simulations, fluctuation theory, molecular motors, tricarboxylic acid cycle, adaptation, biological

INTRODUCTION

Since the time of Boltzmann, it was recognized that living organisms are thermodynamic entities. Lotka (1922a) paraphrased Boltzmann's thinking, "that the fundamental object of contention in the life-struggle, in the evolution of the organic world, is available energy." Lotka went on, "in accord with this observation is the principle that, in the struggle for existence, the advantage must go to those organisms whose energy-capturing devices are most efficient in directing available energy into channels favorable to the preservation of the species." Lotka (1922b) proposed that natural selection is at its most fundamental level a physical principle. Schrödinger (1945) famously expanded on this concept with *What is Life?*, and used the concept of entropy to describe how order, in the form of high energy compounds in the environment, drives organization within organisms. Organisms dissipate that energy into lower forms. The concept of life as a non-equilibrium process has resonated with others as well, including Prigogine who described living organisms as dissipative structures that self-organize in response to large non-equilibrium driving forces (Prigogine, 1978). Abiotic examples of dissipative structures include tornadoes, hurricanes, and convection cells. The non-equilibrium driving forces "pay" for the self-organization that allows the resulting structures to dissipate energy rapidly. In biological systems, energy comes into the system in the form of sunlight or high energy compounds, typically highly reduced carbon compounds, and this energy is dissipated into the environment according to the second law of thermodynamics. In biological systems, some of the energy is harvested to pay for the creation of additional dissipative structures (growth and reproduction), or to create large amounts of stored energy in the form of lower energy byproducts.

The ecologist H. T. Odum was certainly convinced of the role of statistical thermodynamics in systems ecology. Writing in the American Scientist (Odum and Pinkerton, 1955), Odum sought to understand the diverse scale of rates of natural processes, and proposed that each biological system works at an efficiency that allows the maximum efficiency and power, similar to Lotka's concept that the advantage goes to organisms whose metabolism is most efficient at channeling energy for the purpose of reproduction. Odum took natural selection to mean "the persistence of those forms, which can command the greatest useful energy per unit time."

Morowitz also proposed that the far from equilibrium natural environment was responsible for self-organization of biological systems. As a consequence, Morowitz proposed that life was not only a consequence of energy flow in natural systems, but also that it is highly probable. From this perspective, natural selection is a random process, and in the words of Dewar (2005), species "are selected because they are characteristic of each of the overwhelming majority of ways in which energy and matter could flow under the constraints imposed by local energy and mass conservation".

Such concepts have led to the metabolism first hypothesis of the emergence of life on earth (Smith and Morowitz, 2004).

While an excellent collection of discussions of entropy production and self-organization of natural systems has been presented in the literature (Kleidon et al., 2010), for the most part the recognition by physical scientists of the role of thermodynamics as a causal factor in the operation of biological systems stands in stark contrast to the lack of discussion of thermodynamics in the experimental life sciences literature. A major reason for this may be because of the abstract nature of statistical thermodynamics and the lack of tools to model and evaluate the thermodynamic aspects of living systems. After all, since its conceptualization developments in thermodynamics have had mostly to do with equilibrium processes, and biological systems are highly non-equilibrium.

However, in the last 20 years, statistical thermodynamics and fluctuation theorems have allowed for significant progress in understanding non-equilibrium systems. Fluctuation theorems are starting to be used to model biological systems, allowing us to begin to understand how cellular machinery operates. These theorems tell us that there is an important difference between thermodynamic models of macroscopic process and the statistical thermodynamic models of the microscopic processes such as those that make up cells. The second law of thermodynamics describes macroscopic processes and states that the entropy of a spontaneous process never decreases. The second law is silent, however, about the microscopic events that make up the macroscopic process. These microscopic events may be, for instance, sets of coupled reactions that lead to some observable change of state – a different phenotype in the parlance of biology. These microscopic events involve enzyme complexes and coupled reaction pathways in cells, which are not just scaled down versions of beaker-sized laboratory systems. Components of small systems can in fact run in reverse at times. A number of excellent reviews of fluctuation theorems exist in the literature (Harris and Schutz, 2007; Sevick et al., 2008; Seifert, 2012) and we will only give an in-a-nutshell perspective here.

In this report, we will focus on issues and challenges in thermodynamically modeling biological systems of coupled reactions, such as those that occur in metabolism. We will first discuss probability density functions based on Boltzmann probabilities and the relationship to free energy. Closely related to free energy is the concept of entropy. We will discuss different formulations of entropy and their meanings in order to provide a clear overview of entropy production. Finally, fluctuation theorems will be briefly discussed using this conceptual framework. While fluctuation theorems have not yet been used to extensively simulate metabolism, they have great promise, and have been used to examine single molecule dynamics and the dynamics of coupled biochemical reactions on multiple scales. Finally, the application of statistical thermodynamics to model biological reactions that are far from equilibrium is discussed.

THEORETICAL BACKGROUND

Understanding the foundational concepts of modeling thermodynamics is essential for understanding the challenges that the field faces. The mathematical concepts presented in the literature are often too abstract to be readily accessible to those outside the specialty field of statistical thermodynamics. A case in point is that it may seem like the literature contains a zoo of seemingly unrelated statistics all going by the name of entropy. Understanding which entropy is being used is critical for understanding and applying thermodynamic modeling and fluctuation theorems, as will become evident below.

However, a tremendous amount of physical insight into fluctuation theorems and thermodynamic modeling can be obtained if one understands the multinomial distribution function, which is simply a generalization of the common binomial distribution function when more than two outcomes are possible. With regard to reaction kinetics, more than two outcomes are possible when we have more than two interconverting species present. The mathematical form of a multinomial distribution is,

$$\Pr\left(n_1, \ldots, n_\mathrm{m} | N_\mathrm{total}, \theta_1, \ldots, \theta_\mathrm{m}\right) = N_\mathrm{total}! \prod_{\mathrm{objects j}}^{m} \frac{1}{n_\mathrm{j}!} \theta_\mathrm{j}^{n_\mathrm{j}}.$$

The multinomial probability density above is the probability that n_j objects of type j will be present when there are $N_\mathrm{total} = \Sigma n_j$ objects present. In the equation above, θ_j is the probability of object j independent of the other objects. According to frequentist statistics, this probability is simply the long term proportion of the number of object j's that are present, $\theta_j = n_j/N_\mathrm{total}$. The probability density is not simply $\Pr = \Pi_\mathrm{j}\theta_\mathrm{j}^{n_\mathrm{j}}$ because each individual object of type j is indistinguishable from all the other objects of type j. Thus, the probability density has to be corrected for the number of permutations and combinations of each object type, which is accounted for by the factorial terms in the multinomial distribution function.

Now consider a system consisting of three chemical species A, B, and C in aqueous solution in a container of fixed volume. Each of the three species can interconvert to one of the other two species, but the total number of particles is fixed such that $n_\mathrm{A} + n_\mathrm{B} + n_\mathrm{C} = N_\mathrm{total}$. The Boltzmann probability θ_i of species i is related to the Helmholtz free energy of solvation $\Delta\mathcal{A}_\mathrm{i}^0$ by,

$$\theta_\mathrm{i} = \frac{e^{-\Delta\mathcal{A}_\mathrm{i}^0/k_\mathrm{B}T}}{\sum\limits_{\mathrm{species\ j}}^{m} e^{-\Delta\mathcal{A}_\mathrm{j}^0/k_\mathrm{B}T}}. \quad (1)$$

where k_B is Boltzmann's constant and T is the temperature. For simplicity, we will disregard the internal degrees of freedom for each species. In this case, the numerator $e^{-\Delta\mathcal{A}_\mathrm{i}^0/kT}$ is referred to as the molecular partition function, q_i. The denominator is simply a normalization function, usually denoted as $q = \Sigma q_i$, the log of which is the Boltzmann average energy of the system, $-\langle E \rangle_B/k_BT$. Statistically, the distribution of the particles is characterized by the multinomial Boltzmann probability density function,

$$\Pr\left(n_1, \ldots, n_\mathrm{m} | N_\mathrm{total}, \theta_1, \ldots, \theta_\mathrm{m}\right) = N_\mathrm{total}! \prod_{\mathrm{species\ j}}^{m} \frac{1}{n_\mathrm{j}!} \theta_\mathrm{j}^{n_\mathrm{j}}$$

where n_j is the number of particles of species j, and there are N_total particles. In analogy to the macroscopic, the free energy from statistical thermodynamics, an unnormalized mass density for a microscopic state can be defined that is a function of the molecular partition functions q_i instead of the Boltzmann probabilities,

$$\frac{-A\left(n_1,\ \ldots,\ n_{\mathrm{m}}|N_{\mathrm{T}},\ q_1,\ \ldots,\ q_{\mathrm{m}}\right)}{k_{\mathrm{B}}T} = \log\left(N_{\mathrm{total}}!\prod_{\mathrm{j}}^{m}\frac{1}{n_{\mathrm{j}}!}q_{\mathrm{j}}^{n_{\mathrm{j}}}\right) \tag{2}$$

For brevity, we will write $A(n_1,\ \ldots,\ n_{\mathrm{m}}\,|\,N_{\mathrm{r}},\ q_1,\ \ldots,\ q_{\mathrm{m}})$ as $A(\bar{n}|N_{\mathrm{T}},\ \bar{q})$ or simply A. The value A in Eq. 2 is not a free energy because it is not an average over all possible values for each of the n_j. The relationship between A and the probability density of that microscopic state is,

$$-A/k_{\mathrm{B}}T\ = \log\Pr(n_1,\ \ldots,\ n_{\mathrm{m}}|N_{\mathrm{total}},\ \theta_1,\ \ldots,\ \theta_{\mathrm{m}}) + N_{\mathrm{total}}\cdot\log q$$

or equivalently,

$$\log\Pr\left(n_1,\ldots,\ n_{\mathrm{m}}|N_{\mathrm{total}},\ \theta_1,\ldots,\ \theta_m\right) = A/k_{\mathrm{B}}T\ + N_{\mathrm{total}}\cdot\log q$$

Since $\log q = -\langle E\rangle_B/k_BT$, we have the relationship

$$\begin{aligned}-S_{\mathrm{g}} &= A/k_{\mathrm{B}}T - N_{\mathrm{T}}\langle E\rangle_{\mathrm{B}}/k_{\mathrm{B}}T\\ S_{\mathrm{g}} &= -\log\Pr\left(n_1,\ \ldots,\ n_{\mathrm{m}}|N_{\mathrm{T}},\ \theta_1,\ \ldots,\ \theta_{\mathrm{m}}\right)\end{aligned} \tag{3}$$

This function on the right hand side is strictly a log likelihood, not an entropy. However, the average log likelihood is an entropy, and in fact is the Gibbs entropy for a system with a fixed number of total particles,

$$\begin{aligned}S_G &= \sum_{\text{microstates J}}\Pr(J)\log\Pr(J)\\ &= \left\langle A(\bar{n}|N_{\mathrm{T}},\bar{q})\right\rangle - N_{\mathrm{total}}\langle E\rangle_{\mathrm{B}}\end{aligned} \tag{4}$$

where $\Pr(J)$ is shorthand for $\Pr(n_1 = n_1(J), \ldots, n_m(J)|N_{\mathrm{total}}, \theta_1, \ldots, \theta_m)$ and $\left\langle A(\bar{n}|N_{\mathrm{T}},\bar{q})\right\rangle = \mathcal{A}$ is the free energy of the macroscopic state with parameter N_T. Because the Gibbs entropy is an average over microstates, it is the entropy related to macroscopic observations (Jaynes, 1965).

Adding confusion to the definition of entropy is the related microstate relationship,

$$S_{\mathrm{B}} = A/k_{\mathrm{B}}T - N_{\mathrm{total}}\langle E/k_{\mathrm{B}}T\rangle_{\mathrm{U}} \tag{5}$$

where now $\langle E/k_BT\rangle_U$ is the average energy of the microstate under the uniform distribution instead of the Boltzmann distribution. The entropy term is also given by $S = -\Sigma p_j\log p_j$ where again the probabilities $p_j = n_j/N_{\mathrm{total}}$ are from the uniform distribution (Davidson, 1962; Cannon, 2014). The subscript indicates that this is the Boltzmann entropy because it is derived from $\log W$ where W is the multinomial coefficient. This entropy is also sometimes referred to as the configurational entropy (Davidson, 1962). The difference between the Gibbs and Boltzmann entropies of course has to do with intermolecular potentials and microscopic vs. macroscopic perspectives (Jaynes, 1965).

When the total number of particles is not fixed, adjustments need to be made to the equations above. Typically, the adjustment is to remove the normalization of the Boltzmann probabilities in Eq. 1, such that the resulting quantity $e^{-A/k_{\mathrm{B}}T}$ is an unnormalized probability mass function, or an odds of $e^{-A/k_{\mathrm{B}}T}$: 1. The multinomial probability distribution now becomes a multinomial odds distribution, the main difference being that a probability mass function over all of state space sums to 1, while the new multinomial distribution sums to a value >1.

If the total number of particles is allowed to vary due to the system being open, then Eq. 4 gives

$$S_{\mathrm{G}} = \left\langle A - N_{\mathrm{total}}\left(J\right)\log q\right\rangle$$

Notice that this definition is different from one common thermodynamic definition of entropy, which defines entropy as the difference between the free energy and the average energy,

$$\begin{aligned}S &= \langle A\rangle - \langle E\rangle\\ &= \langle A\rangle - \left\langle N_{\mathrm{total}}\left(J\right)\log q\right\rangle\end{aligned}$$

Since we know from the triangle inequality, $\|\log x - \log y\| \geq \|\log x\| - \|\log y\|$, it follows that $S_G \geq S$.

For a set of coupled reactions such as,

$$A \rightleftarrows B \rightleftarrows C$$

a change of the microscopic state from K to J is described by the likelihood ratio,

$$-\Delta S_{\mathrm{g,JK}} = \log\left(\frac{\Pr(J)}{\Pr(K)}\right), \tag{6}$$

or equivalently,

$$\frac{\Pr(J)}{\Pr(K)} = e^{-\Delta S_{\mathrm{g,JK}}} \tag{7}$$

which has the basic mathematical form of a fluctuation theorem, but in this case is an identity due to the definition of S_g in Eq. 3. If we average over all states J and K and the system is at equilibrium,

$$\begin{aligned}\left\langle\frac{\Pr(J)}{\Pr(K)}\right\rangle &= \left\langle e^{-\Delta S_{\mathrm{g,JK}}}\right\rangle\\ &= 1\end{aligned} \tag{8}$$

where the angular brackets denote an equilibrium average. The average value is unity since the log likelihood of Eq. 6 is zero, on average. Relation 8 simply says, that on average, the system returns to equilibrium. While Eq. 7 is exact for microscopic processes, the challenge in employing it to model time-dependent processes is that the core probabilities available for use in Eq. 1 are stationary Boltzmann probabilities, yet if the individual rates of the reactions vary enough in a system of coupled reactions, the core probabilities will not be Boltzmann probabilities, which are based solely on energy levels of the reactants and products. At equilibrium, Eq. 7 can be used for time-dependent probabilities because of detailed balance – Eq. 8. However, away from equilibrium, Eq. 7 no longer holds because detailed balance no longer exists. Instead, the true probabilities will be a function of the entire energy surface of the system, including the reaction barriers. Fluctuation theorems relate the ratio of these time-dependent probabilities to a function that is related to the time-dependent $\Delta S_g(t)$, or if ensemble averages are used, the time-dependent $\Delta S_G(t)$.

For example, at a non-equilibrium steady state the average fluctuations of a system can still be characterized at times without knowing the actual probabilities of each state. Consider the fluctuation away from a steady state J to the new state K with some transition probability. We know that the system will eventually return to the steady state J, we just do not know specifically how. For the most part, a fluctuation away from the steady state will be along the direction of the non-equilibrium driving force. When the system returns to the steady state, an amount of energy will have been dissipated from the system. Note that if the system were to return to the steady state along the same path, no energy would have been dissipated; that is, the average likelihood of returning along the same path is not 1 as in the case for equilibrium (Eq. 8). Thus, fluctuation theorems for non-equilibrium steady state take the form,

$$\Omega = \left\langle \log\left(\frac{\pi_{\mathrm{KJ}}(t)}{\pi_{\mathrm{KJ}}(t)}\right)\right\rangle_{\mathrm{J,K}} \tag{9}$$

where π_{KJ} is the probability of trajectory $J \rightarrow K$, and Ω is related to the dissipation of energy due to the non-equilibrium steady state. For instance, the Evans–Searles fluctuation theory relates the time-dependent probabilities to a trajectory-specific dissipation function, $\Omega(t)$, which is a measure of how far the system is away from detailed balance,

$$\frac{\pi_{\mathrm{KJ}}\left(\Omega(t) = -q_{\mathrm{D}}/k_{\mathrm{B}}T\right)}{\pi_{\mathrm{JK}}\left(\Omega(t) = q_{\mathrm{D}}/k_{\mathrm{B}}T\right)} = e^{-q_{\mathrm{D}}/k_{\mathrm{B}}T} \tag{10}$$

If q_D represents the dissipated energy due to the lack of detailed balance, then the odds of regaining that energy through a reversal of the trajectory are exponentially small. One could even think of the RHS of Eq. 10 as representing the energy of a hypothetical particle (a "dissipation") that has a Boltzmann factor of $e^{-q_{\mathrm{D}}/k_{\mathrm{B}}T}$. Recent developments in fluctuation theories (reviewed by Sevick et al., 2008; Seifert, 2012) in the last two decades have pushed the envelope into the far from equilibrium domain. Many biochemical reactions are in this domain.

ENTROPY PRODUCTION

When the time-dependent flux of material through reactions can be determined, the entropy production rate can be defined in several related ways (Oster et al., 1973; Ge et al., 2006; Ge and Qian, 2010). Using Eq. 6, the microscopic entropy production can be defined for a reaction i in the +direction as,

$$\text{microscopic entropy production rate} = J_{\mathrm{i}+}\Delta S_{\mathrm{g,i}}$$

and the net entropy production through the reaction is $J_{\mathrm{i,net}}\Delta S_{\mathrm{g,i}}$, where $J_{\mathrm{i,net}} = J_{i+} - J_{i-}$. Taking the ratio of the entropy production due to the forward and the reverse reaction, the odds of entropy being produced at reaction i are,

$$\begin{aligned} O\left(\Delta S_{\mathrm{g,i}}\right) &= \frac{J_{\mathrm{i}+} \cdot \Delta S_{\mathrm{g,i}}}{J_{\mathrm{i}-} \cdot \Delta S_{\mathrm{g,i}}} \\ &= \frac{J_{\mathrm{i}+}}{J_{\mathrm{i}-}} \end{aligned} \tag{11}$$

Although the ratio of the forward and reverse flux gives us the odds of thermodynamic entropy production, the ratio itself cannot tell us the value of the thermodynamic entropy change or even if the entropy change is positive or negative; in coupled systems the flux through any specific reaction is not deterministically related to the entropy or free energy change of that reaction. The second law of thermodynamics only tells us that for macroscopic processes, the entropy must always increase; the second law does not address what might be happening on the microscopic level in individual reactions. This is an important aspect of stochastic systems: even though a reaction has a free energy change above zero or equivalently an odds below one, it can still occur given enough time. For example, if a set of coupled reactions has a large enough overall favorable change in free energy, an individual reaction can have a net positive flux even if the reaction free energy is unfavorable. Flux is an emergent property of the entire system. However, as indicated by the fluctuation theorems, the less likely the reaction, the less likely it will have a net flux in the direction of decreasing entropy change.

Several studies have asserted that the relationship between flux and free energy is $\Delta G = -RT\log(J_+/J_-)$. This relationship was originally proposed in discussions of reversible systems and discussed in the context of deterministic kinetics (Beard and Qian, 2007). For coupled, stochastic non-equilibrium reactions, the relationship is strictly speaking an assumption. However, it is reasonable to expect in the vast majority of situations that ΔG and $-RT\log(J_+/J_-)$ are concordant. The relationship can be used to gain insight if used carefully. For instance, Noor et al. (2014) have used the assumption as a framework for evaluating flux statistics at individual reactions. They correctly pointed out that reactions near equilibrium act as kinetic bottlenecks in pathways that are overall far from equilibrium. This is a valid use of the assumption in that reactions at equilibrium in an otherwise nonequilibrium system are those for which the relation is approximately correct even for stochastic systems.

So far the question of how to find the steady states has been left open. A steady state could be determined by the textbook approach of solving the set of differential rate equations. However, for biological systems the required rate parameters are rarely available. In principle, a steady state can be defined based on experimental measurement of all relevant chemical species, which can be used to define the chemical potential of each species. While this task is much easier than determining all the appropriate rate constants, it is still formidable. Yet, significant progress is being made (Bennett et al., 2009).

Alternatively, one can assume that the steady state is one that corresponds to an optimal thermodynamic process. A thermodynamically optimal process is one in which a maximal amount of energy can be extracted from the environment with a minimal amount of dissipation of heat (Sivak and Crooks, 2012). Equivalently, a thermodynamically optimal path is one that requires the least work to maintain the steady state. In either case, the thermodynamically optimal steady state can be found by maximizing a steady state version of Eq. 4 in which the Gibbs entropy S_G in a state space neighborhood Γ measures the probability density of states reachable from an initial state $\mathbb{S}$ due to a series of Z reactions involving a change of state $\delta\mathbb{S}_i$ (Cannon, 2014),

$$S_{\mathrm{g}}\left(\Gamma\left(\mathbb{S}\right)\right) = -\sum_{\mathrm{Rxn}\ \mathrm{i}=1}^{Z} \Pr\left(\mathbb{S}_{\mathrm{i}-1} + \delta\,\mathbb{S}_i\right) \log \Pr\left(\mathbb{S}_{\mathrm{i}-1} + \delta\,\mathbb{S}_{\mathrm{i}}\right) \quad (12)$$

In a system moving toward equilibrium through a trajectory of Z reactions, the state entropy increases as the system stabilizes, and reaches a maximum at equilibrium since equilibrium requires that each respective reaction is equally likely. In a non-equilibrium system, the neighborhood Γ is a reaction path and Eq. 12 is the path entropy described by Dewar, from which the fluctuation theorem, the selection principle of maximum entropy production, and self-organized criticality can be derived (Dewar, 2003). An analogous Gibbs entropy can be defined by averaging $S_g[\Gamma(\mathbb{S})]$ over many trajectories such that $S_G[\Gamma(\mathbb{S})] = \langle S_g[\Gamma(\mathbb{S})]\rangle$. If the entropy change from equilibrium is $\Delta S_{\mathrm{G}}(\Gamma(\mathbb{S}))S_{\mathrm{G}}^{0} - S_{\mathrm{G}}(\Gamma(\mathbb{S}))$, then the rate of production of thermodynamic entropy can then be defined as,

$$\text{thermodynamic entropy production rate} = J_{\mathrm{net}}\left(\Gamma\right)\Delta S_{\mathrm{G}}\left(\Gamma\left(\mathbb{S}\right)\right)$$

While its likely that no individual organism is at the apex of thermodynamic optimality, it is also likely, as discussed in the section "Introduction," that natural selection is at some fundamental level based on filtering out individuals that are thermodynamically inefficient such that too little energy is extracted from the environment or too much of the extracted energy is simply dissipated back to the environment; such a system would not be able to channel sufficient energy into growth to compete against more efficient individuals. In this scenario of natural selection, thermodynamically optimal steady states would serve as useful models.

Applications

Beyond atomistic simulations, the application of statistical thermodynamics and fluctuation theory to biological systems is truly a frontier. To date, applications are mostly in the physics literature and include (but are not limited to) the study of molecular motors, mostly ATP synthase (Andrieux and Gaspard, 2006; Hayashi et al., 2010; Zimmermann and Seifert, 2012), small metabolic networks (Cannon, 2014), bifurcation dynamics of reaction pathways (Xiao et al., 2009), and models of the response of bacteria to changes in the environment (Barato et al., 2014). These examples were chosen to represent a hierarchy of scales in which statistical thermodynamic simulations have been applied to biology. Because the dynamics of each system is represented using different equations, it is not possible to describe in detail the form of the fluctuation theorem used other than to say that all are in some way represented by Eq. 9, except where noted. Details on the theorems are best obtained from the original literature. Below, we briefly summarize the findings for this representative selection from the literature.

SINGLE MOLECULE DYNAMICS OF ATPase F1 ROTARY MOTOR

The F_0F_1–ATP synthase complex is an example of a highly non-equilibrium nanomotor. The rotary motor of F_0F_1–ATP synthase is powered by proton flow across a gradient producing a free energy difference of 10–20 kJ/mol of protons. This free energy difference is significantly greater than the ambient energy at room temperature of about 2.45 kJ/mol. The motor operates over a large range of scales; rate constants for the various processes making up the motor vary over 12 orders of magnitude. Andrieux and Gaspard used fluctuation theory and generating functions to evaluate statistical distributions of mean rotation of the F_1 rotor, the dissipated work, and the probability flux across the system (Andrieux and Gaspard, 2006). The analysis showed that the ATPase motor has a highly non-linear response to chemical fuel: the mean velocity of the F_1 rotor as a function of the thermodynamic driving force is a sigmoid-like curve. Despite the microscopic nature of the motor, the operation of the motor is highly robust in this non-linear regime: successive rotations are statistically correlated and remain essentially unaffected by the fluctuations. Nevertheless, it was shown that the fluctuation theorem held even in the highly non-linear regime.

MULTIPLE MOLECULES: PATHWAY BIFURCATION DYNAMICS OF A CIRCADIAN CLOCK

When multiple reactions are coupled, non-intuitive behavior can result. The Lotka–Voltera oscillator and the Brusselator are famous early examples where feedback or feed-forward interactions control the oscillatory behavior. At the cell level, an important oscillatory phenomenon is the circadian clock of organisms as diverse as fruit flies and fungi. In the circadian clock negative feedback controls, the rate of transcription and translation of specific proteins that in turn dictate the cellular circadian oscillation cycle (Dunlap, 1999).

Using a stochastic thermodynamics approach pioneered by Seifert and colleagues, Xiao et al. (2009) used a chemical Langevin equation to evaluate dynamic bifurcations that occur in the circadian clock. An explicit expression for the mean entropy production in the stationary state was formulated based on available kinetic data. On either side of the bifurcation in the circadian dynamics, the shape of the distribution of the entropy production was similar and highly skewed such that the probability of observing dynamics with negative entropy production was quite small. Thus, like the F1 motor of ATP synthase, the operation of the molecular circadian clock studied by Xiao et al. is robust despite the stochastic nature of small systems.

Although the time dependence of the entropy production in the fluctuation theorem used in this study ultimately came from rate constants, the approach demonstrated that statistical thermodynamic simulations are capable of producing similar bifurcation dynamics as stochastic kinetic simulations. Understanding the entropy production rates of metabolism is important for quantitating the capacity for organisms to adapt to their changing environment, which is discussed next.

CELLULAR INFORMATION PROCESSING AND ADAPTATION

Philosophically, one can adopt either of two opposing perspectives about the relationship between simple biological systems such as bacteria and their environment. One can take the perspective that cells make decisions based on their external environment, which is the most discussed perspective in the literature, or one can take the perspective that the external environment determines cellular response. While the former perspective imbues autonomy to the cell, the latter perspective takes the view that regulation is ultimately a function of the external environment. Who is driving – the cell or the environment? While the former perspective is correct on short time periods such as the diurnal cycles, the latter perspective is more correct on longer time periods over which the cell has adapted and evolved.

Barato et al. (2014), evaluated models of how much information cells can extract from their environment based on their thermodynamic efficiency. Although Barato et al. use the metaphor of learning for the ability to extract information, one is equally justified in using the concept of self-organization. The study found that the degree to which a cell can self-organize in response to the environment is bounded by the thermodynamic entropy production rate. A bacterium in a slowly changing environment dissipates much more energy than it harnesses for the purpose of self-organization. That is, the bacterium, once organized to respond to a particular environment, has a limited ability to further harness energy from the environment for further adaptation.

Although Barato et al. (2014) used quite simple physical models to generate hypotheses, clearly coupling this framework with more extensive thermodynamic models of metabolism has the potential to provide insight into how cells respond internally to changes in environmental driving forces on both short time scales and longer evolutionary time scales. However, modeling efforts will require more sophisticated models of metabolism in order to understand the multitude of paths that cell behavior can take. Next, early efforts that have been taken to expand the application of statistical thermodynamics to more detailed metabolic models are discussed.

DETAILED METABOLIC MODELS

The models and systems discussed above are small systems compared to the metabolism of even the smallest bacterium. Can statistical thermodynamics and fluctuation theories also be applied to more extensive biological systems such as genome-scale models of metabolism? The issue mostly pertains to whether sufficient parameters can be estimated. Large-scale estimates of thermodynamic parameters are available from sources such as the Biochemical Reactions Thermodynamics Database at University of Michigan (Li et al., 2011), the Thermodynamics of Enzyme-Catalyzed Reactions Database at NIST (Goldberg et al., 2004), and the eQuilibrator web server (Flamholz et al., 2012).

We have been developing such an approach and to-date have applied it to relatively small metabolic pathways of various bacteria (Cannon, 2014). In these initial studies, the reactions rates are assumed to be proportional to the thermodynamic driving force of the reaction, which is quantified by a probability of a reaction in a Markov model based on Eq. 7.

Initial studies have focused on the tricarboxylic acid (TCA) cycles of bacteria. These cycles are central to the metabolism of most organisms and may be as close to a universal pathway as there is (Smith and Morowitz, 2004). TCA cycles are capable of consuming acetyl-CoA to either produce high energy compounds necessary for cell function (e.g., ATP, NADPH) or carbon backbones that serve as synthetic precursors for many reactions of secondary metabolism and amino acid and nucleic acid synthesis.

Shown in **Figure 1** is the TCA cycle of *E. coli* and in **Figure 2** is the free energy, energy, and entropy profiles under metabolic conditions observed for exponential growth on glucose (Bennett et al., 2009). The cycle was simulated using statistical thermodynamics formulation of a Markov model based on a local equilibrium assumption (Cannon, 2014). As one proceeds from acetyl-CoA clockwise around the cycle to oxaloacetic acid, the free energy change across the reactions (**Figure 2**) varies smoothly, as one would expect from a maximum entropy perspective (Eq. 12). However, the change for the conversion of oxaloacetate and acetyl-CoA to citrate catalyzed by citrate synthase and the change for the conversion of 2-oxoglutarate to succinyl CoA catalyzed by 2-oxoglutarate dehydrogenase are somewhat abrupt compared to changes at the other reactions of the cycle. The reason for this is that the cofactor concentrations, which serve as boundary conditions, are held fixed at values that prevent the system from relaxing further. As a result, the system is not quite thermodynamically optimal – the entropy defined by Eq. 12 is not quite maximal compared to the value that would be obtained if each reaction was equally likely.

Clearly, information about the thermodynamics of biosynthetic pathways is important for engineering metabolism to overproduce

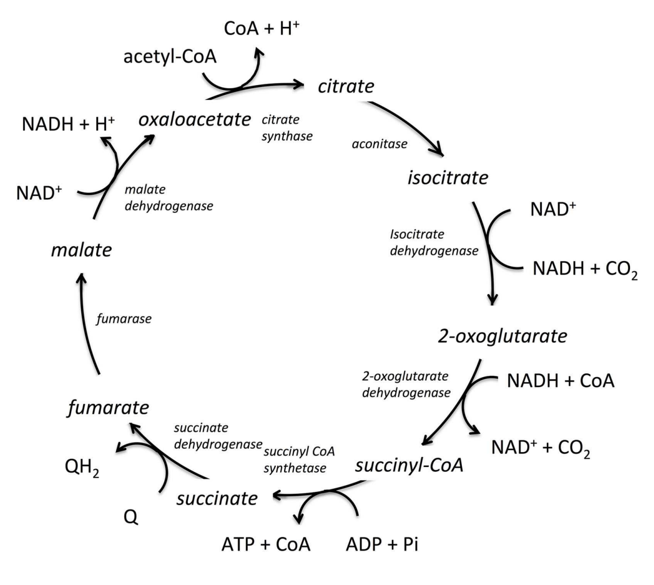


FIGURE 1 | The tricarboxylic acid cycle (TCA) from *E. coli*. The enzymes catalyzing the reactions are shown in italics, the co-factors are shown tangentially to each respective reaction, and the reaction intermediates are shown in line with the cyclic reaction arrows indicating direction of the cycle for *E. coli*. Q and QH$_2$ are electron acceptor/donator pairs and are entry points to the electron transfer chain.

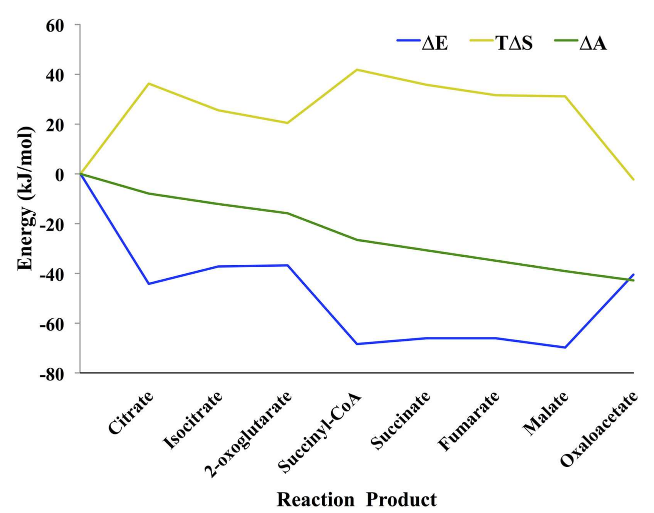


FIGURE 2 | Thermodynamic profile of the TCA cycle from *E. coli* (Cannon, 2014). Eq. 4 was used to calculate the change in entropy ΔS, energy ΔE, and the log of the (unnormalized) mass density $\Delta \mathcal{A}$. Because the probability mass density consists of a combinatorial coefficient that is represented by the entropy term and an energy-based (Boltzmann) probability, there is energy–entropy compensation throughout the cycle. $\Delta \mathcal{A}$ changes smoothly across the reaction pathway indicating that the concentrations of the metabolites are close to optimal, likely because the concentrations were taken from an experimental measurement of *E. coli* metabolite levels.

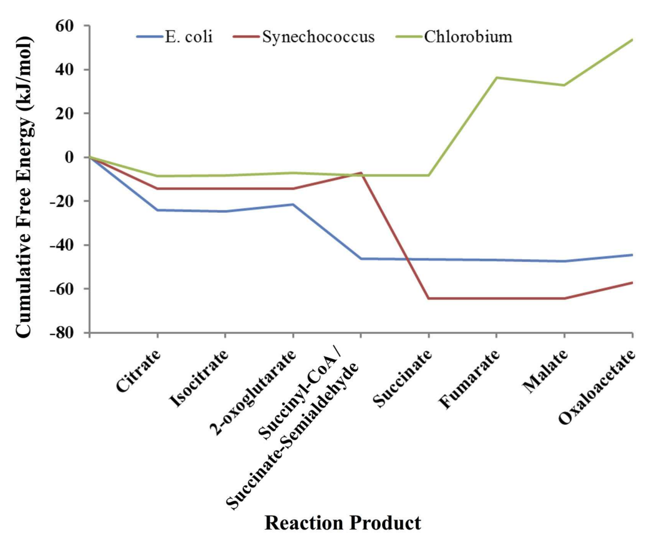


FIGURE 3 | Comparison of the thermodynamic profiles of the TCA cycles of *E. coli, Synechococcus sp. PCC 7002 and Chlorobium tepidum.* The free energy profile of the TCA cycle for each organism reflects its environmental niche (see Discussion).

target compounds such as reduced carbon compounds for biofuels. While much attention has been directed at redirecting carbon flow by knocking out pathways competing for precursors, less attention has been directed at engineering redox pairs such as NADH:NAD$^+$ levels that would thermodynamically drive these reactions. Likewise, much attention has focused on the use of riboswitches to up-regulate the production of enzymes involved in the biosynthesis of target compounds (Wittmann and Suess, 2012), but switching on the catalytic machinery to synthesize a compound is not useful unless the thermodynamics of the pathway are favorable. Modeling metabolic systems thermodynamically would be of enormous value for metabolic engineering.

As an example of the potential use of statistical thermodynamics for both engineering and understand organisms in the context of their natural habitats, we compared three different versions of the TCA cycle used in three very different ecological niches: a typical heterotrophic TCA cycle from *E. coli* involved in extracting energy and biosynthetic precursors from glucose; the cyanobacterial TCA cycle of *Synechococcus* sp. PCC 7002, which is required to produce biosynthetic precursors despite already high levels of ATP from photosynthesis; and the TCA cycle of *Chlorobium tepidum*, a green sulfur bacteria that also must produce biosynthetic precursors in the presence of photosynthesis and simultaneously fix $CO_{2,}$ which it does by running the TCA cycle in the reductive direction. As above, each TCA cycle was simulated using a Markov model based on a local equilibrium assumption. The free energy profiles for these organisms are shown in **Figure 3**. Clearly, each pathway is very different thermodynamically. The cycles for *E. coli* and Synechococcus have similar profiles except for the conversion of 2-oxoglutarate to succinate. In the *E. coli* TCA cycle, this reaction has ATP as a product. *Synechococcus* and other cyanobacteria cannot use the same reaction for converting 2-oxoglutarate to succinate cycle because their cycles must operate in an environment in which ATP concentrations are quite high due to concomitant photosynthesis. Instead, the cyanobacteria use a TCA cycle that employs a ferredoxin coenzyme for this conversion, and thus high levels of ATP do not retard the production of succinate and other carbon compounds that are necessary for growth. The free energy profile of the TCA cycle for *Chlorobium* is very different from both the *E. coli* and *Synechococcus* cycles. Instead of having a highly favorable free energy profile for operation in the oxidative direction (citrate → oxaloacetate), the free energy changes are highly unfavorable. The TCA cycle of *Chlorobium* and other green sulfur bacteria, in fact, runs in the opposite direction (oxaloactetate → citrate), and these organisms use the cycle to fix CO_2 and produce acetyl-CoA. Not only does a thermodynamic model allow us to understand each organism in its environment, but clearly designing an optimal pathway for metabolic engineering using statistical thermodynamics would be very useful.

In comparing the free energy profiles for *E. coli* in **Figures 2** and **3**, it is clear that they differ significantly. In **Figure 2**, the free energy profile changes relatively smoothly as one traverses the cycle, while in **Figure 3** the free energy profile changes abruptly at times. The reason for these differences has to do with the conditions used in the respective simulations. In **Figure 2**, the simulations used the published experimentally measured values for *E. coli* (Bennett et al., 2009). In the latter case, the count of each intermediate in the cycle was initially set to ~20 μm each instead of using the experimental published values for *E. coli* (Bennett et al., 2009), which otherwise might bias the comparison between the three organisms. Although each cycle is materially open in that two carbons come in as acetyl-CoA and carbons leave as CO_2, the total of the number of intermediates is fixed by the stoichiometry of the overall reaction for completion of the cycle. For *E. coli*, the overall stoichiometry is,

$$\begin{aligned}&\text{Acetyl} - \text{CoA} + \text{ADP} + 3\text{NAD}^{+} + \text{P}_{\text{i}} + \text{Q} + 2\text{H}_2\text{O}\\ &\rightleftharpoons \text{CoA} + \text{ATP} + 3\text{NADH} + 2\text{CO}_2 + \text{QH}_2,\end{aligned}$$

where Q and QH_2 represent an oxidized and reduced electron carriers, respectively. Although the cycles are open, the sum of the count of all intermediates will only vary by ±1.

The free energy profiles of the *E. coli* TCA cycle as a function of the total concentration of the intermediates are shown at the top of **Figure 4**. The total concentration values are 1.0-fold, 0.1-fold, 0.01-fold, and 0.001-fold of the values reported by Bennett et al. (2009). If there are only a few total intermediates, then these will be transformed into the metabolites with lowest chemical potentials, which in the case of the *E. coli* TCA cycle are citrate and succinyl CoA. At very low levels of intermediates, the cycle will not operate and citrate and succinyl CoA will simply pool. For the lowest level of intermediates, there will be flux through the entire cycle only over relatively long time periods.

As the total number of metabolic intermediates is raised, the number of citrate and succinyl CoA molecules increase, as shown in **Figure 4** (bottom). Eventually, product builds up as well with a concomitant increase in the free energies of reactions producing citrate and succinyl CoA. Meanwhile, the increase in citrate decreases the free energy for the citrate to isocitrate reaction, and likewise, the increase in succinyl CoA decreases the free energy for the succinyl CoA to succinate reaction.

Eventually, metabolite levels build up to the point where all reactions become equally likely in agreement with Eq. 12. This is thermodynamically the most optimal since the state entropy (Eq. 12) has been maximized with respect to the non-equilibrium boundary conditions.

However, for the cell there is also a thermodynamic penalty to obtain this configuration. In order to handle a greater number of reactants, the enzymatic load on the cell must likewise increase. The self-organized structures needed to dissipate energy rapidly (or store the harvested energy for growth) must be paid for by the non-equilibrium driving forces.

Enzymes catalyzing reactions far from equilibrium will need to increase the least since material flow is unidirectional. This is clearly the case for the enzyme-catalyzed reactions for transformation of oxaloacetate to citrate and 2-oxoglutarate to succinate: as the total metabolite pool increases, the concentrations of the reactants oxaloacetate and 2-oxoglutarate do not change markedly.

If enzymes near equilibrium are expressed at a level just sufficient to catalyze its current load, then increasing the total pool of metabolites may require increased expression of these enzymes. However, these reactions are not likely to remain at equilibrium. This is apparent in **Figure 4** (top) in which the last four enzyme-catalyzed reactions of the TCA cycle transforming succinyl CoA to oxaloacetate, are close to equilibrium when the total pool of metabolites is 0.001-fold of the values reported by Bennett et al. (2009). As the total metabolite pool grows, the reactions do not remain at equilibrium.

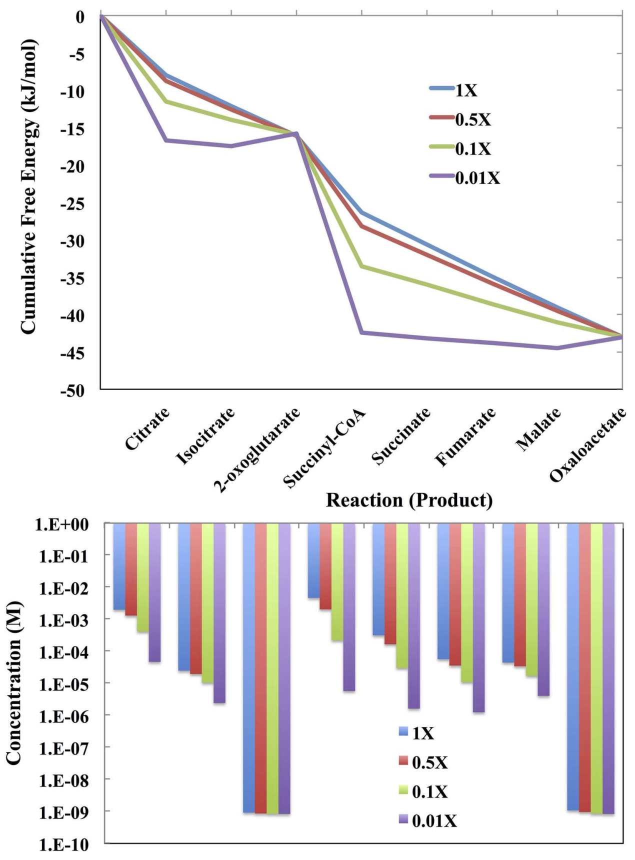


FIGURE 4 | (Top) The cumulative free energy profile of the *E. coli* TCA cycle as a function of the total concentration of the reaction intermediates. Although carbon can enter the cycle as acetyl-coa and leave as CO_2, the total number of intermediates is constrained by the overall reaction (see text). The concentrations used are 1-fold, 0.1-fold, 0.01-fold, and 0.001-fold of those reported by Bennett et al. (2009) for exponential growth on glucose. (Bottom) The distribution of reaction intermediates as a function of total concentration.

When metabolite levels are greater than the respective Michaelis constant (K_M), then enzyme levels need to increase in order to maintain a steady state. This is the situation described by Flamholz et al. (2013). That enzymes catalyzing reactions far from equilibrium do not increase significantly has been experimentally observed; the degree to which enzyme expression will need to increase for reactions near equilibrium will be situation dependent but generally will need to increase with increased flux (Hochachka et al., 1998).

Moreover, if the turnover rates for the enzymes in the pathway differ dramatically, then there must also be a differential level of expression of the enzymes in the pathways. It would make sense for the organism to have high intrinsic enzyme turnover rates for costly enzymes, either those that have many amino acids or require high energy co-factors, such that the thermodynamic cost to the cell can be minimized (Flamholz et al., 2013).

Considering **Figure 4** (top), the data reported by Bennett et al. (2009), implies that the TCA cycle of the laboratory strain of *E. coli* is operating near optimal efficiency with regard to Eq. 12 during exponential growth on glucose. In Lotka's words, "the struggle for existence, the advantage must go to those organisms whose energy-capturing devices are most efficient in directing available energy into channels favorable to the preservation of the species."

How close are biological systems to optimal efficiency? There appear to be situations when this ideal is not achieved. For example, if glycolysis were left unchecked such that each reaction were equally likely thermodynamically, then the large free energy change for conversion of fructose 6-phosphate to fructose 1,6-bisphosphate would result in cellular concentrations of fructose 1,6-bisphosphate several orders of magnitude higher than is observed, which would most likely have detrimental affects on the cell. In fact, the enzyme catalyzing this step is highly regulated to prevent overproduction of fructose 1,6-bisphosphate. The regulation can be regarded as a self-organized and emergent property of the pathway, and one that is necessary for the organism to remain viable. Considering the framework for adaption laid out by Barato et al. (2014), this would imply that for *E. coli* species that are adapted to growth on high levels of glucose, there are very little opportunities for learning alternative ways of regulating this enzyme, or conversely, that the regulatory circuit is evolutionarily stable in this regard.

Future Directions

Determining a rate constant for an enzyme of interest is a straightforward task if the reactant or product has a distinct spectroscopic signature. However, scaling the process up to obtain all of the rate constants necessary for large-scale simulations of metabolism of any specific organism is simply not feasible. Mixing and matching rate constants from orthologous enzymes from different species can result in incorrect energetics, unless one constrains the rate constants to match the equilibrium constant for the same reaction. Moreover, *ad hoc* adjustment of a rate constant to obtain the correct equilibrium constant is likely not better than assuming rates are proportional to the thermodynamic driving force. As a result of the difficulty in obtaining rate constants, constraint-based flux models have been the method of choice for large-scale modeling of biological processes such as metabolism. However, constraint-based methods at best use the thermodynamic constraints to narrow down the solution space. Unfortunately, this limits the predictive power of these approaches.

Several promising and fundamentally sound approaches that include proper thermodynamics have been proposed to move beyond constraint-based flux modeling. One approach is to model systems using mass action kinetics for those reactions for which rate parameters are available, and to use constraint-based flux modeling of other reactions (Chowdhury et al., 2014). In this case, the fluxes modeled using mass action kinetics limit the range of fluxes that are possible for those reactions modeled with constraint-based flux modeling.

A second approach is to use available kinetic parameters where one can, and then infer the remaining parameters based on prior knowledge, including balancing rate parameters to ensure that the correct thermodynamics are obtained (Stanford et al., 2013). An alternative is to reduce the kinetic complexity of the rate equation of each reaction-based analysis of the reaction likelihood as a function of the net flux of the reaction (Canelas et al., 2011). For some reactions, the rate parameters can be eliminated altogether and replaced by the thermodynamic likelihood of the reaction without compromising the fidelity of the model.

Finally, if one knows the reaction directionality, such as from an experimentally based metabolic flux analysis, then a set of feasible metabolite concentrations and reaction free energies can be determined using optimization methods (De Martino et al., 2012). The ability to map out the energy landscape of metabolism could be very powerful and could inform us on whether the conjectures by Lotka, Odum, and others about natural selection discussed in the section "Introduction" are correct. The criteria used by De Martino et al. may actually be too stringent in that the optimization constraints required that the entropy production for each reaction be positive. As indicated in the section "Discussion" around Eq. 11, the second law only requires that the entropy production for the overall macroscopic process be positive. An individual reaction may have a positive flux and also a positive free energy change, but the chance of such an event decreases exponentially with increases in the free energy (Evans and Searles, 1994). The analysis requires the input of flux configurations or reaction directionality. However, this is where fluctuation theories can play a role if they can provide flux values as well.

The use of detailed fluctuation theorems will depend on whether theorems can be developed for non-equilibrium steady states that do not use rate constants and are instead based on chemical potentials and thermodynamic driving forces. If so, then one can set the chemical potentials based (ideally) on metabolomics measurements and carry out large-scale simulations of metabolism that would be identical to kinetic simulations based on rate constants. Experimentally measuring metabolite concentrations is an emerging area of great interest. Key to making the measurements useful for interpretation and modeling is reducing the uncertainty that the measured values reflect *in vivo* concentrations (Noack and Wiechert, 2014).

An alternative statistical thermodynamic approach is to model the process as thermodynamically optimal in which the rates are proportional to the thermodynamic driving force. In a thermodynamically optimal process, the maximum amount of energy is extracted from the environment with a minimal amount of dissipation of heat (Sivak and Crooks, 2012). A model based on this assumption would be roughly consistent with the historical perspectives of the physical basis of biological systems. An analogous approach has been used to analyze metabolomics data, in which the free energies of reactions are minimized with respect to available metabolomics data in order to infer sites of enzyme regulation (Kummel et al., 2006).

As mentioned above, a challenge to using simulations based on statistical thermodynamics is determining accurate standard free energies of reaction or formation of each metabolite. Standard free energies based on group contribution methods are available *en masse* (Jankowski et al., 2008; Noor et al., 2013), but group contribution methods can be inaccurate at times. One must be careful when estimating a standard reaction free energy from group contribution estimates of standard formation free energies in that the errors in estimates are additive; one must ensure when taking the difference between two chemical species that any approximations used for group energies cancel out. The use of electronic structure calculations with an appropriate solvent model is an attractive alternative for determining standard free energies and chemical potentials. Such calculations have been done on a large scale for chlorinated hydrocarbons (Bylaska, 2006) and it is feasible to carry these out for many metabolites. Larger molecules from secondary metabolism, such as those from plants, may present a challenge in that they may have multiple minima that contribute to their free energy of solvation.

ACKNOWLEDGMENTS

This work was supported by the Laboratory Directed Research Program at the Pacific Northwest National Laboratory and the Environmental Molecular Sciences Laboratory (EMSL) at the Pacific Northwest National Laboratory (PNNL). EMSL is a national scientific user facility operated by PNNL for the Office of Biological and Environmental Research at the U. S. Department of Energy. PNNL is operated by Battelle for the U.S. Department of Energy under Contract DE-AC06-76RLO.

REFERENCES

Andrieux, D., and Gaspard, P. (2006). Fluctuation theorems and the nonequilibrium thermodynamics of molecular motors. *Phys. Rev. E* 74, 011906. doi:10.1103/PhysRevE.74.011906

Barato, A. C., Hartich, D., and Seifert, U. (2014). Efficiency of cellular information processing. *New J. Phys.* 16, 103024. doi:10.1088/1367-2630/16/10/103024

Beard, D. A., and Qian, H. (2007). Relationship between thermodynamic driving force and one-way fluxes in reversible processes. *PLoS ONE* 2:e144. doi:10.1371/journal.pone.0000144

Bennett, B. D., Kimball, E. H., Gao, M., Osterhout, R., Van Dien, S. J., and Rabinowitz, J. D. (2009). Absolute metabolite concentrations and implied enzyme active site occupancy in *Escherichia coli*. *Nat. Chem. Biol.* 5, 593–599. doi:10.1038/nchembio.186

Bylaska, E. J. (2006). Estimating the thermodynamics and kinetics of chlorinated hydrocarbon degradation. *Theor. Chem. Acc.* 116, 281–296. doi:10.1007/s00214-005-0042-8

Canelas, A. B., Ras, C., Ten Pierick, A., Van Gulik, W. M., and Heijnen, J. J. (2011). An in vivo data-driven framework for classification and quantification of enzyme kinetics and determination of apparent thermodynamic data. *Metab. Eng.* 13, 294–306. doi:10.1016/j.ymben.2011.02.005

Cannon, W. R. (2014). Simulating metabolism with statistical thermodynamics. *PLoS ONE* 9:e103582. doi:10.1371/journal.pone.0103582

Chowdhury, A., Zomorrodi, A. R., and Maranas, C. D. (2014). k-OptForce: integrating kinetics with flux balance analysis for strain design. *PLoS Comput. Biol.* 10:e1003487. doi:10.1371/journal.pcbi.1003487

Davidson, N. (1962). *Statistical Mechanics*. New York, NY: McGraw.

De Martino, D., Figliuzzi, M., De Martino, A., and Marinari, E. (2012). A scalable algorithm to explore the Gibbs energy landscape of genome-scale metabolic networks. *PLoS Comput. Biol.* 8:e1002562. doi:10.1371/journal.pcbi.1002562

Dewar, R. (2003). Information theory explanation of the fluctuation theorem, maximum entropy production and self-organized criticality in non-equilibrium stationary states. *J. Phys. A Math. Gen.* 36, 631–641. doi:10.1088/0305-4470/36/3/303

Dewar, R. (2005). "Maximum entropy production and non-equilibrium statistical mechanics," in *Non-equilibrium Thermodynamics and the Production of Entropy*, eds A. Kleidon and R. Lorenz (Berlin: Springer), 41–55.

Dunlap, J. C. (1999). Molecular bases for circadian clocks. *Cell* 96, 271–290. doi:10.1016/S0092-8674(00)80566-8

Evans, D. J., and Searles, D. J. (1994). Equilibrium microstates which generate 2nd law violating steady-states. *Phys. Rev. E* 50, 1645–1648. doi:10.1103/PhysRevE.50.1645

Flamholz, A., Noor, E., Bar-Even, A., Liebermeister, W., and Milo, R. (2013). Glycolytic strategy as a tradeoff between energy yield and protein cost. *Proc. Natl. Acad. Sci. U.S.A.* 110, 10039–10044. doi:10.1073/pnas.1215283110

Flamholz, A., Noor, E., Bar-Even, A., and Milo, R. (2012). eQuilibrator – the biochemical thermodynamics calculator. *Nucleic Acids Res.* 40, D770–D775. doi:10.1093/nar/gkr874

Ge, H., Jiang, D. Q., and Qian, M. (2006). Reversibility and entropy production of inhomogeneous Markov chains. *J. Appl. Probab.* 43, 1028–1043. doi:10.1239/jap/1165505205

Ge, H., and Qian, H. (2010). Physical origins of entropy production, free energy dissipation, and their mathematical representations. *Phys. Rev. E* 81, 051133. doi:10.1103/PhysRevE.81.051133

Goldberg, R. N., Tewari, Y. B., and Bhat, T. N. (2004). Thermodynamics of enzyme-catalyzed reactions – a database for quantitative biochemistry. *Bioinformatics* 20, 2874–2877. doi:10.1093/bioinformatics/bth314

Harris, R. J., and Schutz, G. M. (2007). Fluctuation theorems for stochastic dynamics. *J. Stat. Mech.* 2007, 07020. doi:10.1088/1742-5468/2007/07/P07020

Hayashi, K., Ueno, H., Iino, R., and Noji, H. (2010). Fluctuation theorem applied to F1-ATPase. *Biophys. J.* 98, 633A–633A. doi:10.1016/j.bpj.2009.12.3466

Hochachka, P. W., Mcclelland, G. B., Burness, G. P., Staples, J. F., and Suarez, R. K. (1998). Integrating metabolic pathway fluxes with gene-to-enzyme expression rates. *Comp. Biochem. Physiol. B* 120, 17–26. doi:10.1016/S0305-0491(98)00019-4

Jankowski, M. D., Henry, C. S., Broadbelt, L. J., and Hatzimanikatis, V. (2008). Group contribution method for thermodynamic analysis of complex metabolic networks. *Biophys. J.* 95, 1487–1499. doi:10.1529/biophysj.107.124784

Jaynes, E. T. (1965). Gibbs vs. Boltzmann entropies. *Am. J. Phys.* 33, 391–398. doi:10.1119/1.1971557

Kleidon, A., Malhi, Y., and Cox, P. M. (2010). Maximum entropy production in environmental and ecological systems. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 365, 1297–1302. doi:10.1098/rstb.2010.0018

Kummel, A., Panke, S., and Heinemann, M. (2006). Putative regulatory sites unraveled by network-embedded thermodynamic analysis of metabolome data. *Mol. Syst. Biol.* 2, 2006.0034. doi:10.1038/msb4100074

Li, X., Wu, F., Qi, F., and Beard, D. A. (2011). A database of thermodynamic properties of the reactions of glycolysis, the tricarboxylic acid cycle, and the pentose phosphate pathway. *Database (Oxford)* 2011, bar005. doi:10.1093/database/bar005

Lotka, A. J. (1922a). Contribution to the energetics of evolution. *Proc. Natl. Acad. Sci. U.S.A.* 8, 147–151. doi:10.1073/pnas.8.6.147

Lotka, A. J. (1922b). Natural selection as a physical principle. *Proc. Natl. Acad. Sci. U.S.A.* 8, 151–154. doi:10.1073/pnas.8.6.151

Noack, S., and Wiechert, W. (2014). Quantitative metabolomics: a phantom? *Trends Biotechnol.* 32, 238–244. doi:10.1016/j.tibtech.2014.03.006

Noor, E., Bar-Even, A., Flamholz, A., Reznik, E., Liebermeister, W., and Milo, R. (2014). Pathway thermodynamics highlights kinetic obstacles in central metabolism. *PLoS Comput. Biol.* 10:e1003483. doi:10.1371/journal.pcbi.1003483

Noor, E., Haraldsdottir, H. S., Milo, R., and Fleming, R. M. (2013). Consistent estimation of Gibbs energy using component contributions. *PLoS Comput. Biol.* 9:e1003098. doi:10.1371/journal.pcbi.1003098

Odum, H. T., and Pinkerton, R. T. (1955). Time's speed regulator: the optimum efficiency for maximum power output in physical and biological systems. *Am. Sci.* 43, 331–343.

Oster, G. F., Perelson, A. S., and Katchals, A. (1973). Network thermodynamics – dynamic modeling of biophysical systems. *Q. Rev. Biophys.* 6, 1–134. doi:10.1017/S0033583500000081

Prigogine, I. (1978). Time, structure, and fluctuations. *Science* 201, 777–785. doi:10.1126/science.201.4358.777

Schrödinger, E. (1945). *What is Life? The Physical Aspect of the Living Cell*. Cambridge: The University Press.

Seifert, U. (2012). Stochastic thermodynamics, fluctuation theorems and molecular machines. *Rep. Prog. Phys.* 75, 126001. doi:10.1088/0034-4885/75/12/126001

Sevick, E. M., Prabhakar, R., Williams, S. R., and Searles, D. J. (2008). Fluctuation theorems. *Annu. Rev. Phys. Chem.* 59, 603–633. doi:10.1146/annurev.physchem.58.032806.104555

Sivak, D. A., and Crooks, G. E. (2012). Thermodynamic metrics and optimal paths. *Phys. Rev. Lett.* 108, 190602. doi:10.1103/PhysRevLett.108.190602

Smith, E., and Morowitz, H. J. (2004). Universality in intermediary metabolism. *Proc. Natl. Acad. Sci. U.S.A.* 101, 13168–13173. doi:10.1073/pnas.0404922101

Stanford, N. J., Lubitz, T., Smallbone, K., Klipp, E., Mendes, P., and Liebermeister, W. (2013). Systematic construction of kinetic models from genome-scale metabolic networks. *PLoS ONE* 8:e79195. doi:10.1371/journal.pone.0079195

Wittmann, A., and Suess, B. (2012). Engineered riboswitches: expanding researchers' toolbox with synthetic RNA regulators. *FEBS Lett.* 586, 2076–2083. doi:10.1016/j.febslet.2012.02.038

Xiao, T. J., Hou, Z. H., and Xin, H. W. (2009). Stochastic thermodynamics in mesoscopic chemical oscillation systems. *J. Phys. Chem. B* 113, 9316–9320. doi:10.1021/jp901610x

Zimmermann, E., and Seifert, U. (2012). Efficiencies of a molecular motor: a generic hybrid model applied to the F-1-ATPase. *New J. Phys.* 14, 20. doi:10.1088/1367-2630/14/10/103023

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.

Received: 09 September 2014; accepted: 27 October 2014; published online: 26 November 2014.

Citation: Cannon WR (2014) Concepts, challenges, and successes in modeling thermodynamics of metabolism. Front. Bioeng. Biotechnol. ***2****:53. doi: 10.3389/fbioe.2014.00053*

This article was submitted to Systems Biology, a section of the journal Frontiers in Bioengineering and Biotechnology.

frontiers in
BIOENGINEERING AND BIOTECHNOLOGY

REVIEW ARTICLE
published: 03 December 2014
doi: 10.3389/fbioe.2014.00062

Metabolic network discovery by top-down and bottom-up approaches and paths for reconciliation

Tunahan Çakır[1]* *and Mohammad Jafar Khatibipour*[1,2]

[1] Computational Systems Biology Group, Department of Bioengineering, Gebze Technical University (formerly known as Gebze Institute of Technology), Gebze, Turkey

[2] Department of Chemical Engineering, Gebze Technical University (formerly known as Gebze Institute of Technology), Gebze, Turkey

Edited by:
Daniel Machado, University of Minho, Portugal

Reviewed by:
Pierre Millard, University of Manchester, UK
Corrado Priami, University of Trento and COSBI, Italy
Huub C. J. Hoefsloot, University of Amsterdam, Netherlands

Correspondence:
Tunahan Çakır, Department of Bioengineering, Gebze Technical University, Gebze 41400, Kocaeli, Turkey
e-mail: tcakir@gyte.edu.tr

The primary focus in the network-centric analysis of cellular metabolism by systems biology approaches is to identify the active metabolic network for the condition of interest. Two major approaches are available for the discovery of the condition-specific metabolic networks. One approach starts from genome-scale metabolic networks, which cover all possible reactions known to occur in the related organism in a condition-independent manner, and applies methods such as the optimization-based Flux-Balance Analysis to elucidate the active network. The other approach starts from the condition-specific metabolome data, and processes the data with statistical or optimization-based methods to extract information content of the data such that the active network is inferred. These approaches, termed bottom-up and top-down, respectively, are currently employed independently. However, considering that both approaches have the same goal, they can both benefit from each other paving the way for the novel integrative analysis methods of metabolome data- and flux-analysis approaches in the post-genomic era. This study reviews the strengths of constraint-based analysis and network inference methods reported in the metabolic systems biology field; then elaborates on the potential paths to reconcile the two approaches to shed better light on how the metabolism functions.

Keywords: constraint-based models, metabolic network inference, active metabolic state, metabolome, network biology, reverse engineering, flux-balance analysis

INTRODUCTION

Metabolic network is the outmost layer of cellular activity from the genome. The genome of a cell is a comprehensive and condensed information base, defining a boundary for the biochemical capacity of the cell. The processing of genetic information passes through several layers of fabrication and regulation before reaching their end products. This is from information to the function, from genotype to phenotype. Metabolic enzymes count for a significant percentage of the end products of genes, and their activity sets the physiology of the cell. Since metabolic network activity is the major representative of cell functionality, it is of great importance to gain as much knowledge as possible on the active metabolic network at a specific cellular state.

Systems-based approach to molecular biology has contributed to an increased knowledge of metabolic pathways for an increasing number of organisms, and led to almost complete metabolic networks for a number of major organisms, from yeast to human. Such static networks are available in a condition-independent manner through web-based databases such as KEGG or MetaCyc (Altman et al., 2013), or reconstructed in a format suitable for simulation by several researchers at genome scale (Oberhardt et al., 2009; Kim et al., 2012). There are several mathematical approaches to process such networks to come up with condition-specific networks, the most common one being the Flux-Balance Analysis (FBA) framework (Orth et al., 2010). This is a bottom-up direction toward the active network since already-known "parts," interactions, are used as inputs (Bruggeman and Westerhoff, 2007; Petranovic and Nielsen, 2008).

In parallel to the developments on the knowledge of metabolic networks, techniques to measure metabolite levels at high throughput, termed metabolomics, have arisen (Kell, 2004; Dunn et al., 2005). Quantitative or semi-quantitative metabolome data, although one of the most challenging compared to other omic sciences, have come a long way in a decade, from the detection and quantification of about 50 metabolites (Devantier et al., 2005) to more than 1000 metabolites (Psychogios et al., 2011). Metabolome data are a snapshot of the condition-specific status of the investigated organisms. Reverse-engineering metabolome data to discover the underlying network structure is the goal behind metabolic network inference approaches (Srividhya et al., 2007; Çakır et al., 2009). The information content of metabolome data is revealed by processing it with correlation or optimization-based methods (Weckwerth et al., 2004; Hendrickx et al., 2011; Öksüz et al., 2013). Such an approach to discover metabolic network structure is termed top-down approach since the parts, interactions, are not known *a priori*, and predicted from the whole set of available biomolecules (Bruggeman and Westerhoff, 2007; Petranovic and Nielsen, 2008).

In this review, we will cover the basic developments in bottom-up and top-down approaches to discover active metabolic network, and then ponder over the possible ways of reconciling these two approaches for a better prediction of active

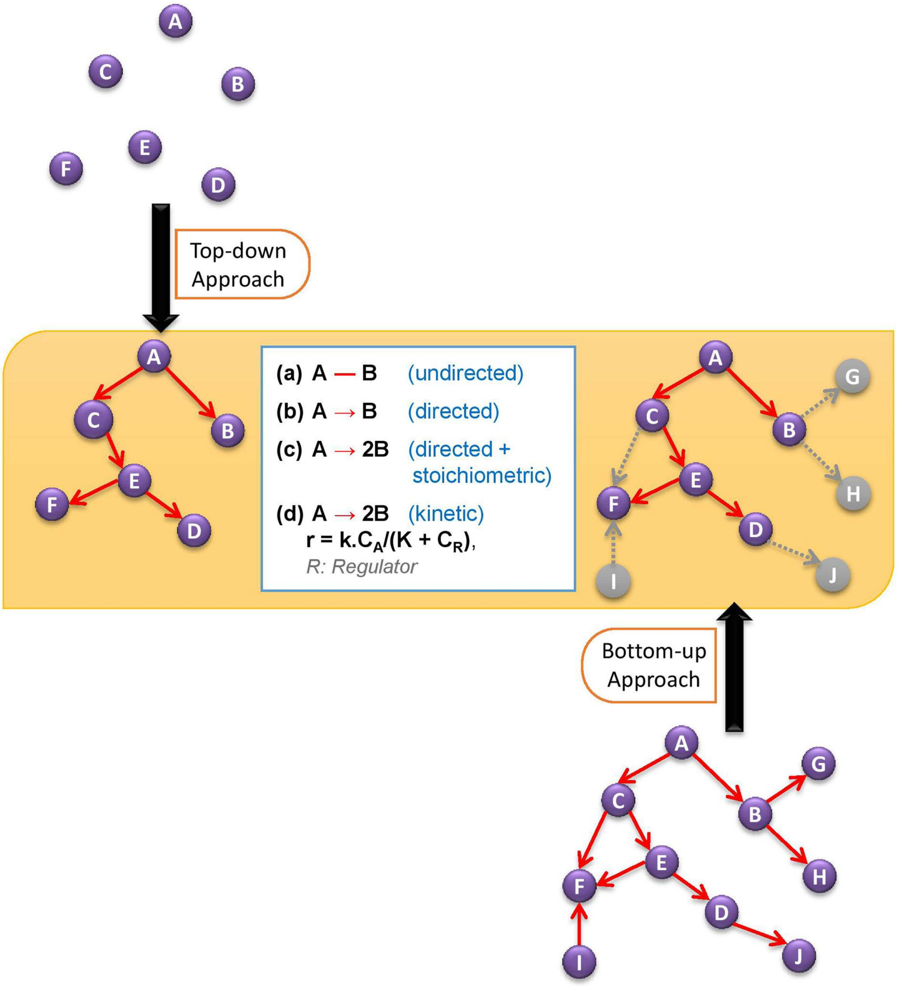


FIGURE 1 | Comparative demonstration of bottom-up and top-down approaches to discover active metabolic network. The white box in the figure defines different levels of network structure information.

network structure. **Figure 1** illustrates the two alternative network discovery approaches.

BOTTOM-UP APPROACHES TO DISCOVER CONDITION-SPECIFIC METABOLIC NETWORKS

Different methods and algorithms have been used for the discovery and characterization of active metabolic networks at different states of cells and culture environments. In the bottom-up approach, everything starts from an already available network of biochemical transformations that cover all possible scenarios in the distribution of metabolic fluxes, and sets an upper bound for the existence of reactions in the active metabolic network. Such a network is termed a static metabolic network. A static metabolic network can be provided either by a previously reconstructed genome-scale stoichiometric model or by a collection of all reactions whose existence in the organism of interest has been certified in literature and databases. Most popular among such databases are KEGG (Kanehisa et al., 2014), MetaCyc (Caspi et al., 2014), and Reactome (Croft et al., 2014). Other efforts with more curated databases such as Rhea (Alcántara et al., 2012) and MetRxn (Kumar et al., 2012) are also available. A genome-scale stoichiometric model is reconstructed based on the annotation of all genes in the genome of one organism to their end products and then to the corresponding reactions, leading to a list of gene-protein-reaction rules (Thiele and Palsson, 2010). In this way, the minimum information content of a genome-scale model is (i) a list of reactions, and (ii) a list of gene-protein-reaction rules. The presence of gene-protein-reaction rules in stoichiometric models has enabled the opportunity for transcriptome and proteome data to be incorporated into the discovery methods of active metabolic networks (Blazier and Papin, 2012).

Given a genome-scale reaction network, the aim is to find the active reaction network at a specific condition or for a specific cell type in a multicellular organism (**Box 1**). The core of all such discovery approaches is a stoichiometric matrix. Each row of the stoichiometric matrix represents a metabolite and each column stands for a reaction, the corresponding element being the stoichiometric coefficient of that metabolite in that reaction. The relationship

Box 1 | Different levels of Metabolic Network Structure Information.

Our understanding of an active metabolic network can be sorted into several stages of information.

(i) At the lowest level of information, we want to know what the structure of the network is, representing it with an undirected (or directed, if the reversibility information is available) graph in which each node stands for a metabolite and each edge stands for a biochemical transformation. Alternative to the retrieval from the metabolic reaction databases, the structure of the network – both directed and undirected – can also be estimated to some extent by analyzing and reverse engineering the metabolome data without the use of *a priori* database information on the reactions.

(ii) At a higher level, the information on the stoichiometry of reactions can be incorporated, leading to a directed stoichiometric biochemical network.

(iii) Having the stoichiometric structure of the network, we can characterize the metabolic state in more detail by quantifying the metabolic fluxes. In most cases, rather than a unique flux distribution, constraints are set on flux values to shrink the solution space. Such modeling approaches are known as Constraint-Based Modeling. This level of understanding the active metabolic network (structure + flux distribution) has been the area of focus in the research community for more than a decade. In most cases, the information provided at this level has been satisfactory for engineering research to design more efficient cell factories, and also, recently, for medical research to distinguish significant differences between healthy and disease states.

(iv) There are, however, certain limitations at the above level although it provides a network activity structure weighted with fluxes. The dynamic behavior of the system cannot be captured, and the predictability power of such models is hampered mainly because they are not considering the role of regulatory mechanisms in controlling the rate of biochemical reactions. In some cases, the regulation of reaction rates plays such a dominant role that it would be hard to make any prediction by just considering the flux-based network activity structure. Here come the kinetic models into the picture, which take enzymatic regulations and metabolite concentrations into account for a dynamic and better prediction of network structure.

between the reaction rates in the network and the dynamic change in the concentration of metabolites is represented as given below:

$$\frac{d\mathbf{C}}{dt} = \mathbf{S} \times \mathbf{v} \tag{1}$$

where **S** is the stoichiometric matrix, **C** is the vector of intracellular metabolite concentrations, and **v** is a column vector of metabolic reaction rates (fluxes) to be determined. Under the assumption of steady state, the concentration of each intracellular metabolite is not going to change with time, meaning the sum of rate of reactions producing that metabolite is equivalent to the sum of rate of reactions consuming that metabolite (metabolic fluxes around each metabolite are balanced). This is represented mathematically as follows:

$$\mathbf{S} \times \mathbf{v} = \mathbf{0} \tag{2}$$

This is an algebraic system of linear equations with all fluxes being zero as a trivial solution. In order to escape from the trivial solution, the value of at least one of the fluxes must be set to a non-zero value, that flux usually being an exchange flux between the intracellular and extracellular environment since the experimental measurement of exchange fluxes is relatively easier. The system is almost always underdetermined with a large solution space, mainly because of the existence of branch points in the metabolic network. There are both experimental and computational approaches to estimate a condition-specific network for such a system.

The experimental approach is based on stable-isotope (mostly 13C carbon) labeling of the major carbon source, and then tracing the propagation of the labeled carbon atoms down to protein-bound amino acids at isotopic steady state by using mass spectrometry or NMR spectroscopy (Wiechert et al., 2001; Sauer, 2006; Mueller and Heinzle, 2013). The qualitative isotopic labeling information is then used as an input to two alternative methods. In one method, termed isotopomer modeling, a total flux distribution is estimated based on the experimental labeling results through a computationally demanding non-linear optimization formulation, which employs global iterative fitting and statistical analysis (Wiechert et al., 2001; Antoniewicz et al., 2007). The other 13C-labeling assisted method is based on the estimation of the local ratios of fluxes emerging from a branch point (Sauer, 2006; Zamboni et al., 2009) rather than the absolute quantification of all fluxes. These experimental flux split ratios can be used to shrink the solution space of Eq. 2 in a complementary flux calculation, leading to the discovery of a condition-specific network (Schuetz et al., 2007; Tarlak et al., 2014). Softwares are available for the rather sophisticated calculation of experimental fluxes (or flux ratios) from carbon labeling data for both methods (Zamboni et al., 2005; Quek et al., 2009; Weitzel et al., 2013). A new trend in this area is to collect data at the non-stationary phase of isotopic labeling rather than at the isotopic steady state, which was shown to be more informative in terms of predicting the flux-weighted active metabolic network structure (Schaub et al., 2008; Young et al., 2008; Wiechert and Nöh, 2013). Works on the tracing of intracellular metabolites rather than only 10–15 protein-bound amino acids have also appeared due to the higher coverage of metabolic pathways despite the inherent experimental difficulties in terms of higher turnover rates as well as stability issues (Van Winden et al., 2005; Toya et al., 2007; Millard et al., 2014).

The computational approach for the discovery of condition-specific metabolic network based on Eq. 2 is known as constraint-based modeling. Constraint-based modeling methods aim to shrink the solution space of the equation as much as possible by putting relevant constraints on the system. The most common method, FBA, treats the problem in Eq. 2 as an optimization problem and linear programing is applied to solve it. The stoichiometry of metabolic reactions (stoichiometric matrix), reaction directionality information, a physiologically relevant objective function, and the value of at least one of the exchange fluxes are all that are required for FBA to return a condition-specific flux distribution. The flux distribution returned by FBA is not necessarily unique, and there may be a variety of flux distributions all leading to the same optimum value of the objective function. Therefore, Flux Variability Analysis (FVA) must be used together with FBA, to determine the variability, if any, on each metabolic flux in regard to the condition of interest (Mahadevan and Schilling, 2003; Müller and Bockmayr, 2013). The maximization of biomass production has been successfully applied as a reliable objective function for FBA to predict flux distributions in a variety of microorganisms (Varma and Palsson, 1994; Feist and Palsson, 2010). In some studies, it has been hypothesized that one objective function alone may not capture the metabolic behavior of the cell comprehensively. Therefore, multi-objective optimization platforms have been designed and utilized to come up with more specific flux distributions. Several modified versions of FBA including parsimonious FBA, pFBA (Lewis et al., 2010), and flexible-optimality FBA, flexoFBA (Tarlak et al., 2014), have been developed in this manner. On the other hand, some research groups have developed methods based on the availability of additional omics data, which are discussed below. For a thorough review of a number of FBA-derived flux calculation methods, the readers are referred to Lewis et al. (2012).

CONSTRAINTS BASED ON TRANSCRIPTOME OR PROTEOME DATA

The rate of an enzymatic reaction inside the cell is a function of several different factors, such as the concentration of substrates, products, and regulators of the enzyme and also the amount of available active enzyme for that reaction. Among these factors, the concentration of active enzymes can be related to the activity of genes through layers of transcription, translation, and post-translational modifications. Transcriptome data are much more accessible and comprehensive compared to the other omics data. Several different research groups have developed different strategies to incorporate transcriptome data into constraint-based models. The idea behind this is that the amount of mRNAs (gene activities) may be correlated with the concentration of active enzymes, and hence this can be utilized to provide additional constraints on metabolic fluxes. At the bottom line, if an enzyme coding gene is not transcribed at steady state, the corresponding reaction should be inactive at that steady state, if there is no other enzyme catalyzing that reaction. This idea was first used by Akesson et al. to set the flux values to zero for those reactions whose corresponding genes were expressed at low levels (Åkesson et al., 2004). More sophisticated and structured versions of this approach appeared later, under the names of GIMME (Becker and Palsson, 2008) and iMAT (Shlomi et al., 2008). These approaches classify some reactions as inactive reactions based on the low expression levels of their associated genes. Then, they employ a computational framework which minimizes the contradiction between the classification and an active physiological flux distribution since some of these classifications may render the flux state unrealistic (such as zero growth rate). Several other alternative methods appeared recently to incorporate transcriptome data into the prediction of active metabolic network and flux distribution. In an interesting study, for example, mRNA levels from transcriptome data were used as weights for the corresponding reactions to predict a flux distribution without using a conventional objective function such as the maximization of biomass growth (Lee et al., 2012). A recent study (Machado and Herrgård, 2014) evaluated these methods systematically for the prediction of flux distributions, and the results were compared to that of parsimonious FBA as a reference method that does not consider the transcriptome data. In general, none of the methods could significantly improve the results of pFBA and none of them outperformed the others for the tested cases (*S. cerevisiae* and *E. coli*). Instead of the prediction of flux distributions, these methods, however, may significantly help in the discovery of active metabolic networks in context/tissue-specific cells and in the conditions where a relevant objective function cannot be hypothesized.

Transcriptome data are not necessarily correlated with the rate of corresponding reactions. Inconsistency between mRNA levels and reaction rates is a result of influence of several other factors in the regulation of enzymatic reactions. Therefore, if proteome data are available, it can be used instead of transcriptome data as a better representative for the concentration of active enzymes since proteome is hierarchically closer to the enzyme states than transcriptome data. The methods that are developed to integrate transcriptome data with the FBA method can all be used for the purpose of integrating proteome data. For example, GIMMEp (Bordbar et al., 2012) is the proteome equivalent version of GIMME. Some of such integrative methods were primarily tested with proteome data. INIT (Agren et al., 2012), for example, was developed by using proteome abundance data from Human Protein Atlas database. However, it was shown that utilizing proteome data instead of transcriptome data could not improve the prediction of flux distributions for the tested cases (*S. cerevisiae* and *E. coli*) (Machado and Herrgård, 2014). In a study which used metabolome and proteome data in the flux calculation method, on the other hand, even the use of only proteome data were shown to improve the results compared to the traditional FBA (see the next section for more details) (Yizhak et al., 2010).

Substrate concentrations, the concentration of enzyme regulators, the turn over number of the catalyzing enzyme, and the concentration of the active enzyme are all playing significant roles in the determination of reaction rates, and among them only the concentration of the active enzyme may be represented by the corresponding protein or mRNA concentration. Translated proteins are not necessarily active enzymes, and they may need to undergo post-translational modifications (e.g., phosphorylation/acetylation) to become capable of catalyzing the reactions. This is one of the main reasons behind inconsistency between protein levels and reaction rates. On the other hand, the turn over number (catalytic power) of one enzyme may differ by several orders of magnitude from the turn over number of another enzyme (Hoppe, 2012). It means that although the concentration of one enzyme may be much less than the others in the network, the reaction catalyzed by that enzyme can proceed much faster than others. According to this fact, the use of the absolute concentrations of proteins or mRNAs to constrain reaction rates does not seem promising. However, the turn over number of one enzyme in an individual is an intrinsic parameter of the enzyme that does not change from one condition to another except by effective mutations that rarely occur. Because of this, the relative levels of proteins or mRNAs can be utilized to overcome the problem of big differences in turn over numbers. One steady state with available data on flux values and protein/mRNA levels can be taken as the reference state, and then the relative/differential levels of proteins/mRNAs to the reference state can be used to predict the flux distributions at the new conditions. Based on this approach, algorithms have been developed to incorporate relative/differential transcriptome data into metabolic-flux analysis, among which are MADE (Jensen and Papin, 2011) and GX-FBA (Navid and Almaas, 2012). One other main reason for the inconsistency between protein levels and reaction rates is the distribution of flux control among different layers from genotype to phenotype. Metabolic fluxes can be regulated hierarchically (through gene expression levels) or metabolically (through metabolic interactions) (Daran-Lapujade et al., 2007; Postmus et al., 2008; Nikerel et al., 2012; Chubukov et al., 2013). Use of transcriptome or proteome data will not be helpful if the metabolic fluxes are controlled at the metabolic level.

CONSTRAINTS BASED ON METABOLOME DATA

One approach to find more specific and physiologically relevant flux distributions is to provide additional constraints by specifying the directionality of reversible reactions. This can be done by taking Gibbs free energies of metabolites into consideration. The Gibbs free energy change of a reversible biochemical transformation (one reaction or a series of reactions) determines the direction of that transformation and its departure from reversibility. The earlier studies assumed standard conditions (all metabolite concentrations were assumed to be 1 M), and did not explicitly consider metabolite concentrations in the calculation of Gibbs energy changes of reactions due to the scarcity of metabolome data (Henry et al., 2006). Recent studies, however, take the concentration of metabolites into account, when available, to perform thermodynamic-based metabolic-flux analysis, leading to more reliable predictions (Hoppe et al., 2007; Bennett et al., 2009; Soh and Hatzimanikatis, 2010; Hamilton et al., 2013).

Extracellular metabolome data can be used to constrain genome-scale metabolic models for the calculation of intracellular flux distributions by simply constraining the secretion and uptake rates of extracellular metabolites based on such data (Çakır et al., 2007; Mo et al., 2009). In a different approach, Michaelis–Menten-based kinetics was used for the estimation of reaction rates for the reactions for which appropriate intracellular metabolome (and proteome) data are available (Yizhak et al., 2010). The FBA framework was designed in such a way that the calculated fluxes are as consistent as possible with the kinetically derived reaction rates, if available. The simultaneous use of metabolome and proteome data for this purpose significantly improved the results. The use of metabolome data alone also resulted in better predictions than the traditional FBA. In a recent study, a kinetic platform was established based on Michaelis–Menten equation to bridge gene expression levels, metabolite concentrations and metabolic fluxes without requiring the knowledge of kinetic parameters (Zelezniak et al., 2014). They could show that changes in metabolite concentrations relative to a reference steady state can be predicted by their formulation that includes information on network connectivity in addition to differential mRNA expression levels. All those works utilizing kinetic information demonstrate the necessity of dynamic models for a more comprehensive analysis of metabolic networks.

Kinetic models of biochemical reactions not only provide a rational platform for omics data – especially metabolomics – to be incorporated in the estimation of metabolic fluxes but also they enable the prediction and study of the dynamics of metabolic networks far beyond the steady state (**Box 1**). Such models were only possible for small-scale metabolic networks until recently (Teusink et al., 2000; Chassagnole et al., 2002), since, they require detailed information on the enzyme kinetics of each individual reaction. Estimation of kinetic parameters is a major obstacle in the applicability of dynamic modeling of metabolic networks. New platforms and algorithms were established to circumvent this problem so that the estimation of explicit kinetic parameters is not a prerequisite to study the dynamic capacity and behavior of the system (Link et al., 2014). Approximative kinetic models (lin-log, power-law, mass-action) on the other hand, try to fit a standard rate expression formula to all reactions of the network to increase the range of their applicability to larger networks (Visser et al., 2004; Sorribas et al., 2007). Thanks to approximative kinetics, attempts to reconstruct large-scale kinetic metabolic models with more than 100 reactions were recently presented (Smallbone et al., 2010; Chakrabarti et al., 2013; Stanford et al., 2013), but their prediction power is limited to the conditions adequately close to the corresponding steady state.

As a better alternative to approximative kinetics, an approach was established and utilized based on the concept of parametric Jacobian, which covers the behavior of all possible kinetic models that are consistent with an experimentally observed operating point (Steuer et al., 2006). This approach provides an opportunity to detect and analyze bifurcation characteristics of the metabolic network without the need for explicit determination of kinetic parameters. Ensemble modeling of metabolic networks (Tran et al., 2008) is an elegant idea for large-scale kinetic modeling of biochemical reaction networks. In this method, each enzymatic reaction is broken down to its elementary reactions that all follow mass-action kinetics. An ensemble of thermodynamically consistent kinetic models with different dynamic behavior that all converge to a reference steady state is collected with the help of intracellular metabolome data. This ensemble is then filtered by the results of perturbation experiments to filter out inconsistent models from the ensemble and to increase the predictability of remaining models. The approach was successfully applied, among others, to construct kinetic models of *E. coli* (Khodayari et al., 2014) and cancer metabolisms (Khazaei et al., 2012), leading to promising flux predictions.

TOP-DOWN APPROACHES TO DISCOVER CONDITION-SPECIFIC METABOLIC NETWORKS

Time series of metabolite concentrations in response to a perturbation, and also replicates of metabolome data at a specific steady state, both implicitly contain information on the structure of active metabolic network. Reverse engineering of these data to infer the condition-specific metabolic network without necessarily prior knowledge on the genome of the organism and its static metabolic network is an alternative to all bottom-up approaches that are based on the availability of a large-scale stoichiometric model of the organism. Although promising, less attention has been paid to these top-down approaches compared to bottom-ups mainly because of the technical obstacles in gathering reliable metabolome data in large scale. This limitation will be removed with future advancements in the detection and quantification of intracellular metabolites such as higher coverage and temporal resolution. At this stage, however, several research groups have established algorithms and methods for reverse engineering of metabolic networks by using either time series or steady-state replicates of metabolite concentrations (Crampin et al., 2004; Chou and Voit, 2009; Hendrickx et al., 2011; Lecca and Priami, 2013).

NETWORK DISCOVERY BASED ON TIME-SERIES DATA

The use of time-series metabolite concentration data to predict the underlying network connectivity information first appeared in the literature about two decades ago. Time-lagged correlations combined with a projection technique called multidimensional scaling were shown to construct the structure of generic biochemical networks with few nodes (Arkin and Ross, 1995). Correlation between time-series profiles of metabolites, with the consideration of the delay in the influence of one metabolite on the next, is the basis of the time-lagged correlation method for the inference of metabolic networks. The approach, called correlation metric construction, was later experimentally verified *in vitro* by inferring the first steps of glycolytic pathway in a 14-metabolite system (Arkin et al., 1997). Modified versions of the approach appeared later (Samoilov et al., 2001; Lecca et al., 2012). In the latter, metabolic pathway of an anticancer drug was deduced from the time-lagged correlations of corresponding metabolite concentration measurements. The modification introduced by the former work was recently improved by using mutual information similarity score rather than simple linear correlation (Villaverde et al., 2014). The authors compared their method, called MIDER, with several other methods by applying it to different types of cellular networks, including *in vitro* glycolytic pathway data. The approach outperformed the other methods.

Another method to reconstitute a network using time-series data is based on perturbation experiments around steady state. The initial curve of concentration changes of metabolites in response to a pulse change on the concentration of a metabolite is processed with the method of zero initial slopes (Vance et al., 2002). The method successfully inferred the structure of glycolysis based on *in vitro* experimental data (Torralba et al., 2003). Performance comparison of the method with the correlation metric construction approach was later provided based on *in silico* data of *S. cerevisiae* and *E. coli* central metabolic networks (Hendrickx et al., 2011). An approach based also on perturbation experiments, but with a different formulation aiming to calculate Jacobian matrix from time derivatives of concentration data, was first applied to gene networks (Schmidt et al., 2005). A modified version of the approach recently used *in vivo* metabolite concentration measurements from tomato seedlings to reconstruct quercetin glycosylation pathway (Astola et al., 2011).

Apart from such model-free structure identification methods, model-based methods use time-series metabolite concentration data not only to identify network structure but also to estimate proper model parameters such as rate constants of kinetic expressions (Chou and Voit, 2009). Majority of these approaches use power-law (also called S-system) formulation (Savageau and Voit, 1987) to approximate reaction kinetics. An approach, for example, used S-system modeling with a multi-objective optimization by simultaneously minimizing the number of interactions and the error in the fitting (Liu and Wang, 2008). They applied their method to major metabolites involved in ethanol fermentation. An earlier work analyzed a small three-metabolite network of phospholipid metabolism by combining S-system modeling and an evolutionary modeling method, genetic programing (Ando et al., 2002). Later, a new representation of S-system approach, called S-trees, was combined with genetic programing to reverse-engineer yeast fermentation pathway in a more efficient manner by using *in silico* time-series concentration data of five metabolites (Cho et al., 2006). In a sophisticated approach, others used symbolic regression based on genetic programing to infer both the structure and the model of yeast glycolytic oscillations from *in silico* data (Schmidt et al., 2011). Their use of acylic graph encoding rather than tree-based encoding together with symbolic regression approach ensured the identification of parsimonious (sparse) models. Rather than S-system formulation, mass-action kinetics can also be used to infer pathway connectivity and reaction mechanism (Srividhya et al., 2007). This minimizes the computational burden on the algorithm since only rate constants are to be estimated as parameters in the mass-action formulation. The authors tested their method with real time course experimental metabolome data of *Lactococcus lactis* glycolysis. A graphical user interface was later made available by the same group to ease the inference of kinetics and network architecture from dynamic data of biochemical pathways (Mourão et al., 2011). Genetic programing was also combined with mass-action kinetics in an algorithm, which ensures the estimation of biochemically more plausible models (Gormley et al., 2013). The small phospholipid network of (Ando et al., 2002) was inferred in a more compact way by this algorithm.

NETWORK DISCOVERY BASED ON STEADY-STATE DATA

The use of steady-state metabolome data to infer metabolic network structure has also drawn attention in the last decade. The biological variability in the metabolism of the organisms at around steady state is a known phenomenon due to slight variations in the enzyme levels or due to slight natural or environment-induced fluctuations within cellular processes. Slight variations in the steady-state measurements of metabolite levels can be informative on the network structure (Steuer et al., 2003; Camacho et al., 2005; Çakır et al., 2009). The most common approach here is to use the similarity measures such as Pearson correlation to assign edges between metabolites. One should note that such correlations are not necessarily strong among neighboring metabolites whereas there could be strong correlations among distant metabolites in the network (Camacho et al., 2005). In a comprehensive study, different alternative similarity measures (linear vs. non-linear, and full vs. partial) were applied to *in silico* metabolome data belonging to two microorganisms to systematically analyze method performances (Çakır et al., 2009). The results revealed no clear superiority between linear (Pearson correlation) and non-linear (mutual information) similarity measures. The best performing method was identified as nth order partial Pearson correlation, known also as graphical Gaussian modeling. Graphical Gaussian modeling was also applied to metabolome data from blood serum samples to reconstruct human fatty acid metabolism (Krumsiek et al., 2011). Others (Nemenman et al., 2007) analyzed *in silico* metabolome data of red blood cell metabolism by ARACNE approach (Margolin et al., 2006), which is based on pruning mutual information scores. An elegant improvement on ARACNE based reverse engineering of metabolic profiling data was suggested later (Bandaru et al., 2011). The approach puts a constraint on the possible metabolic transformations to satisfy the mass conservation between the connected metabolites. Synthetic data covering up to about 200 metabolites were generated to test the approach. One issue in such similarity-based approaches is that only pairwise interactions are aimed to be found. However, a metabolic reaction can involve more than two metabolites. Based on this reasoning, an attempt to also deduce triple interactions by using ternary mutual information was suggested (Diệp et al., 2011). Analysis of synthetic yeast glycolysis data and red blood cell data showed the success of this approach in capturing higher order interactions.

A different approach to discover active metabolic networks from steady-state data is based on Lyapunov equation. In Eq. 1, the rate vector, **v**, is a complex non-linear function of concentrations, **C**. For systems around steady state, the equation can be expressed in terms of Jacobian matrix, **J**, by the help of linear approximation:

$$\frac{d\mathbf{X}}{dt} \approx \mathbf{JX} \tag{3}$$

with $\mathbf{X} = \mathbf{C} - \mathbf{C_s}$, and $\mathbf{C_s}$ shows the steady-state metabolite concentrations. Jacobian matrix stores detailed information on the structure of the underlying network; such as the directionality of interaction, strength of interaction, and regulation type of interaction. For small fluctuations around steady state, the right-hand side of Eq. 3 becomes zero, and the left-hand side can be expressed in such a way that a link between the covariance matrix of metabolome data, $\mathbf{\Gamma}$, and Jacobian matrix is provided. The details of the derivation are given elsewhere (Van Kampen, 1992; Steuer et al., 2003).

$$\mathbf{J\Gamma} + \mathbf{\Gamma J}^{\mathrm{T}} = -2\mathbf{D} \tag{4}$$

D in the equation shows the extent of fluctuations. Eq. 4, known as Lyapunov equation, can be used to infer metabolic network structure since it provides a link between the data-based covariance matrix and network connectivity stored in **J**. Reverse-engineering metabolome data by using the Lyapunov equation was first discussed via a hypothetical three-metabolite system (Steuer et al., 2003). A recent work provided a theoretical analysis on the use of the Lyapunov equation to infer network structure from steady-state metabolome data (Öksüz et al., 2013). The authors used a rearranged version of the Lyapunov equation:

$$\mathbf{Aj} = 2\mathbf{d} \tag{5}$$

Here, **j** and **d** are vectorized versions of **J** and **D** matrices. **A** is a matrix based on the covariance of data. In that work, directed networks were inferred from *in silico* metabolome data of *S. cerevisiae* glycolysis, *E. coli* central carbon metabolism, and brain glycolysis by solving Eq. 5 for **j** using a genetic-algorithm based formulation. In the optimization formulation, the dual objective function was simultaneous maximization of the sparse structure and minimization of the residual norm of the equation. When compared to the inference results based on nth order partial Pearson correlation, a much higher prediction accuracy was reported. One other advantage of the optimization-based approach is the fact that Eq. 5 infers a directed network whereas correlation-based approaches cannot predict directions of interactions. The Lyapunov equation was recently used to infer differential changes in Jacobian matrix rather than the inference of network structure by predicting Jacobian matrix itself (Sun and Weckwerth, 2012; Kügler and Yang, 2014; Nägele et al., 2014).

PATHS TO RECONCILE BOTTOM-UP AND TOP-DOWN METABOLIC NETWORK DISCOVERY APPROACHES

Previous sections reviewed bottom-up and top-down metabolic network discovery approaches from literature. Top-down approaches are dependent on intracellular metabolome data, and there are bottom-up approaches, which aim to use omics data as additional constraints. The simultaneous use of both approaches to discover better condition-specific networks has not been a focus in the scientific community. Here, we will elaborate on the ways to reconcile these two approaches when intracellular metabolome data of a condition in question are available.

All model-based top-down approaches using time-series data also infer a Jacobian matrix of the model. Many other top-down approaches are based on correlations between metabolites. There is a significant relationship between the correlation strengths and the strengths of interactions implied by Jacobian entries (Çakır et al., 2009). Therefore, correlation strengths or Jacobian-interaction strengths of the inferred edges can be used as edge scores in the bottom-up constraint-based modeling approaches as additional constraints for a better identification of the active metabolic network as follows: all inferred edges in a top-down approach based on metabolome data are ranked with respect to their edge scores. Afterward, cut-off values for high- and low-scores are determined. If a high-score edge also appears in the corresponding static genome-scale stoichiometric model, that reaction is assigned a high weight. If a high edge-score does not have a corresponding connection in the genome-scale model, this could imply a novel or a regulatory interaction. As it is known, genome-scale metabolic models do not account for regulatory interactions of metabolites with enzymes, however, top-down approaches do not have this limitation since they are purely data-based. If the edge-score is low, the corresponding reaction in the stoichiometric model is assigned a low weight. Similarly, if the top-down approach assigns no edge between two metabolites, which are linked with a reaction in the stoichiometric model, such reactions are also assigned low weight. All other reactions can be assigned with a medium-weight. Then, a mixed-integer programing based optimization framework can be used with Eq. 2 such that the resulting condition-specific flux distribution is as consistent as possible with the edge scores, including maximum possible number of high-weight reactions and minimum possible number of low-weight reactions as active. Thereby, the strength of top-down predictions can be used for better bottom-up flux predictions.

Use of transcriptome or proteome data as constraints in metabolic-flux calculations resulted in several alternative methods such as GIMME, iMAT, and INIT. These approaches remove reactions from the static metabolic reaction set if the controlling gene or protein is not active. However, a recent work comparing all these methods could not identify a method with clear superiority over the parsimonious FBA (Machado and Herrgård, 2014). This approach can be combined with edge scores (inferred Jacobian-interaction strength or calculated correlation strength) information to yield better network identification. GIMME-like approaches remove reactions from the model, this means also removal of metabolites. Two different approaches can be used: (i) removed reactions whose main substrates and products show high edge scores must be retained in the reaction set, implying an active edge (ii) reactions whose main substrates and products show very low and insignificant correlations must be candidates to be removed from the reaction set, implying an inactive edge if their removal does not hamper the objective function. Such a flux calculation powered by the top-down inference of network edges can lead to a more refined network.

One reconciliation approach will be the integrative use of flux-balance equation (Eq. 2) and rearranged Lyapunov equation (Eq. 5). Flux-balance equation was widely used in the last two decades because of its simplicity, requiring only the stoichiometric coefficients of reactions, and few measurement constraints. The rearranged Lyapunov equation bears a similar simplicity since it is only based on the covariances of metabolome measurements. The only major issue, as it is the case in flux-balance equation, is a proper choice of objective function to solve the equation. Since both **J** and **v**, the unknowns in both equations, represent the active network structure, the coupled use of these two equations can be beneficial from two different aspects: (i) a better flux distribution can be found thanks to the metabolome-based constraint provided by Eq. 5, (ii) the information stored in stoichiometric matrix, since it will reveal all possible non-interacting pairs, will provide a constraint to get a better estimate of Jacobian matrix by setting edge scores of some pairs to zero.

An approach getting popular to construct genome-scale kinetic models is ensemble modeling. This modeling approach constructs kinetic models from an ensemble of models, and filters the inconsistent models out by using the results of perturbation experiments (Tran et al., 2008; Khodayari et al., 2014). On the other hand, a number of methods infer networks from time-series data by using a model-based approach. The output of such methods is both the network structure and the dynamic kinetic model with estimated parameters (Srividhya et al., 2007; Liu and Wang, 2008). A number of alternative models are scanned in these methods to infer the most suitable one. Therefore, the strengths of model-based network inference and ensemble-based kinetic model reconstruction can be combined to yield better frameworks.

In summary, both bottom-up and top-down discovery of metabolic networks have come a long way in the last 20 years, providing the scientific community with a number of computational methods, as reviewed in this review. Considering the improvements that are being experienced both on the coverage and precision of metabolome data, the coming decade will witness an exponential increase in the number of metabolome datasets, similar to what was experienced with transcriptome data in the last decade. This review aimed at drawing attention to this point, as ways to reconcile the two major metabolic network discovery approaches will gain increasing importance.

ACKNOWLEDGMENTS

The financial support by TUBITAK, The Scientific and Technological Research Council of Turkey, through a career grant (Project Code: 110M464) is gratefully acknowledged.

REFERENCES

Agren, R., Bordel, S., Mardinoglu, A., Pornputtapong, N., Nookaew, I., and Nielsen, J. (2012). Reconstruction of genome-scale active metabolic networks for 69 human cell types and 16 cancer types using INIT. *PLoS Comput. Biol.* 8:e1002518. doi:10.1371/journal.pcbi.1002518

Åkesson, M., Förster, J., and Nielsen, J. (2004). Integration of gene expression data into genome-scale metabolic models. *Metab. Eng.* 6, 285–293. doi:10.1016/j.ymben.2003.12.002

Alcántara, R., Axelsen, K. B., Morgat, A., Belda, E., Coudert, E., Bridge, A., et al. (2012). Rhea – a manually curated resource of biochemical reactions. *Nucleic Acids Res.* 40, D754–D760. doi:10.1093/nar/gkr1126

Altman, T., Travers, M., Kothari, A., Caspi, R., and Karp, P. D. (2013). A systematic comparison of the MetaCyc and KEGG pathway databases. *BMC Bioinformatics* 14:112. doi:10.1186/1471-2105-14-112

Ando, S., Sakamoto, E., and Iba, H. (2002). Evolutionary modeling and inference of gene network. *Inf Sci.* 145, 237–259. doi:10.1016/S0020-0255(02)00235-9

Antoniewicz, M. R., Kelleher, J. K., and Stephanopoulos, G. (2007). Elementary metabolite units (EMU): a novel framework for modeling isotopic distributions. *Metab. Eng.* 9, 68–86. doi:10.1016/j.ymben.2006.09.001

Arkin, A., and Ross, J. (1995). Statistical construction of chemical reaction mechanisms from measured time-series. *J. Phys. Chem.* 99, 970–979. doi:10.1021/j100003a020

Arkin, A., Shen, P., and Ross, J. (1997). A test case of correlation metric construction of a reaction pathway from measurements. *Science* 277, 1275–1279. doi:10.1126/science.277.5330.1275

Astola, L., Groenenboom, M., Roldan, V. G., Van Eeuwijk, F., Hall, R. D., Bovy, A., et al. (2011). "Metabolic pathway inference from time series data: a non iterative approach," in *Pattern Recognition in Bioinformatics*, eds M. Loog, L. Wessels, M. J. T. Reinders, and D. de Ridder (Berlin: Springer), 97–108.

Bandaru, P., Bansal, M., and Nemenman, I. (2011). Mass conservation and inference of metabolic networks from high-throughput mass spectrometry data. *J. Comput. Biol.* 18, 147–154. doi:10.1089/cmb.2010.0222

Becker, S. A., and Palsson, B. O. (2008). Context-specific metabolic networks are consistent with experiments. *PLoS Comput. Biol.* 4:e1000082. doi:10.1371/journal.pcbi.1000082

Bennett, B. D., Kimball, E. H., Gao, M., Osterhout, R., Van Dien, S. J., and Rabinowitz, J. D. (2009). Absolute metabolite concentrations and implied enzyme active site occupancy in *Escherichia coli*. *Nat. Chem. Biol.* 5, 593–599. doi:10.1038/nchembio.186

Blazier, A. S., and Papin, J. A. (2012). Integration of expression data in genome-scale metabolic network reconstructions. *Front. Physiol.* 3:299. doi:10.3389/fphys.2012.00299

Bordbar, A., Mo, M. L., Nakayasu, E. S., Schrimpe-Rutledge, A. C., Kim, Y.-M., Metz, T. O., et al. (2012). Model-driven multi-omic data analysis elucidates metabolic immunomodulators of macrophage activation. *Mol. Syst. Biol.* 8, 558. doi:10.1038/msb.2012.21

Bruggeman, F. J., and Westerhoff, H. V. (2007). The nature of systems biology. *Trends Microbiol.* 15, 45–50. doi:10.1016/j.tim.2006.11.003

Çakır, T., Efe, Ç, Dikicioglu, D., Hortaçsu, A., Kırdar, B., and Oliver, S. G. (2007). Flux balance analysis of a genome-scale yeast model constrained by exometabolomic data allows metabolic system identification of genetically different strains. *Biotechnol. Prog* 23, 320–326. doi:10.1021/bp060272r

Çakır, T., Hendriks, M. M., Westerhuis, J. A., and Smilde, A. K. (2009). Metabolic network discovery through reverse engineering of metabolome data. *Metabolomics* 5, 318–329. doi:10.1007/s11306-009-0156-4

Camacho, D., de la Fuente, A., and Mendes, P. (2005). The origin of correlations in metabolomics data. *Metabolomics* 1, 53–63. doi:10.1007/s11306-005-1107-3

Caspi, R., Altman, T., Billington, R., Dreher, K., Foerster, H., Fulcher, C. A., et al. (2014). The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases. *Nucleic Acids Res.* 42, D459–D471. doi:10.1093/nar/gkt1103

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

Chassagnole, C., Noisommit-Rizzi, N., Schmid, J. W., Mauch, K., and Reuss, M. (2002). Dynamic modeling of the central carbon metabolism of *Escherichia coli. Biotechnol. Bioeng.* 79, 53–73. doi:10.1002/bit.10288

Cho, D.-Y., Cho, K.-H., and Zhang, B.-T. (2006). Identification of biochemical networks by S-tree based genetic programming. *Bioinformatics* 22, 1631–1640. doi:10.1093/bioinformatics/btl122

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

Chubukov, V., Uhr, M., Le Chat, L., Kleijn, R. J., Jules, M., Link, H., et al. (2013). Transcriptional regulation is insufficient to explain substrate-induced flux changes in *Bacillus subtilis. Mol. Syst. Biol.* 9, 709. doi:10.1038/msb.2013.66

Crampin, E. J., Schnell, S., and McSharry, P. E. (2004). Mathematical and computational techniques to deduce complex biochemical reaction mechanisms. *Prog. Biophys. Mol. Biol.* 86, 77–112. doi:10.1016/j.pbiomolbio.2004.04.002

Croft, D., Mundo, A. F., Haw, R., Milacic, M., Weiser, J., Wu, G., et al. (2014). The Reactome pathway knowledgebase. *Nucleic Acids Res.* 42, D472–D477. doi:10.1093/nar/gkt1102

Daran-Lapujade, P., Rossell, S., van Gulik, W. M., Luttik, M. A. H., de Groot, M. J. L., Slijper, M., et al. (2007). The fluxes through glycolytic enzymes in *Saccharomyces cerevisiae* are predominantly regulated at posttranscriptional levels. *Proc. Natl. Acad. Sci. U.S.A.* 104, 15753–15758. doi:10.1073/pnas.0707476104

Devantier, R., Scheithauer, B., Villas-Bôas, S. G., Pedersen, S., and Olsson, L. (2005). Metabolite profiling for analysis of yeast stress response during very high gravity ethanol fermentations. *Biotechnol. Bioeng.* 90, 703–714. doi:10.1002/bit.20457

Diệp, N. Q., Hoan, P. T., Bảo, H. T., Hùng, T. Đ, and Thắng, P. Q. (2011). Computational reconstruction of metabolic networks from high-throughput profiling data. *J. Comput. Sci. Cybern* 27, 23–35. doi:10.15625/1813-9663/27/1/460

Dunn, W. B., Bailey, N. J., and Johnson, H. E. (2005). Measuring the metabolome: current analytical technologies. *Analyst* 130, 606–625. doi:10.1039/b418288j

Feist, A. M., and Palsson, B. O. (2010). The biomass objective function. *Curr. Opin. Microbiol.* 13, 344–349. doi:10.1016/j.mib.2010.03.003

Gormley, P., Li, K., Wolkenhauer, O., Irwin, G. W., and Du, D. (2013). Reverse engineering of biochemical reaction networks using co-evolution with eng-genes. *Cogn. Comput.* 5, 106–118. doi:10.1007/s12559-012-9159-y

Hamilton, J. J., Dwivedi, V., and Reed, J. L. (2013). Quantitative assessment of thermodynamic constraints on the solution space of genome-scale metabolic models. *Biophys. J.* 105, 512–522. doi:10.1016/j.bpj.2013.06.011

Hendrickx, D. M., Hendriks, M. M. W. B., Eilers, P. H. C., Smilde, A. K., and Hoefsloot, H. C. J. (2011). Reverse engineering of metabolic networks, a critical assessment. *Mol. Biosyst.* 7, 511–520. doi:10.1039/c0mb00083c

Henry, C. S., Jankowski, M. D., Broadbelt, L. J., and Hatzimanikatis, V. (2006). Genome-scale thermodynamic analysis of *Escherichia coli* metabolism. *Biophys. J.* 90, 1453–1461. doi:10.1529/biophysj.105.071720

Hoppe, A. (2012). What mRNA abundances can tell us about metabolism. *Metabolites* 2, 614–631. doi:10.3390/metabo2030614

Hoppe, A., Hoffmann, S., and Holzhütter, H.-G. (2007). Including metabolite concentrations into flux balance analysis: thermodynamic realizability as a constraint on flux distributions in metabolic networks. *BMC Syst. Biol.* 1:23. doi:10.1186/1752-0509-1-23

Jensen, P. A., and Papin, J. A. (2011). Functional integration of a metabolic network model and expression data without arbitrary thresholding. *Bioinformatics* 27, 541–547. doi:10.1093/bioinformatics/btq702

Kanehisa, M., Goto, S., Sato, Y., Kawashima, M., Furumichi, M., and Tanabe, M. (2014). Data, information, knowledge and principle: back to metabolism in KEGG. *Nucleic Acids Res.* 42, D199–D205. doi:10.1093/nar/gkt1076

Kell, D. B. (2004). Metabolomics and systems biology: making sense of the soup. *Curr. Opin. Microbiol.* 7, 296–307. doi:10.1016/j.mib.2004.04.012

Khazaei, T., McGuigan, A., and Mahadevan, R. (2012). Ensemble modeling of cancer metabolism. *Front. Physiol.* 3:135. doi:10.3389/fphys.2012.00135

Khodayari, A., Zomorrodi, A. R., Liao, J. C., and Maranas, C. D. (2014). A kinetic model of *Escherichia coli* core metabolism satisfying multiple sets of mutant flux data. *Metab. Eng.* 25, 50–62. doi:10.1016/j.ymben.2014.05.014

Kim, T. Y., Sohn, S. B., Kim, Y. B., Kim, W. J., and Lee, S. Y. (2012). Recent advances in reconstruction and applications of genome-scale metabolic models. *Curr. Opin. Biotechnol.* 23, 617–623. doi:10.1016/j.copbio.2011.10.007

Krumsiek, J., Suhre, K., Illig, T., Adamski, J., and Theis, F. J. (2011). Gaussian graphical modeling reconstructs pathway reactions from high-throughput metabolomics data. *BMC Syst. Biol.* 5:21. doi:10.1186/1752-0509-5-21

Kügler, P., and Yang, W. (2014). Identification of alterations in the Jacobian of biochemical reaction networks from steady state covariance data at two conditions. *J. Math. Biol.* 68, 1757–1783. doi:10.1007/s00285-013-0685-3

Kumar, A., Suthers, P. F., and Maranas, C. D. (2012). MetRxn: a knowledgebase of metabolites and reactions spanning metabolic models and databases. *BMC Bioinformatics* 13:6. doi:10.1186/1471-2105-13-6

Lecca, P., Morpurgo, D., Fantaccini, G., Casagrande, A., and Priami, C. (2012). Inferring biochemical reaction pathways: the case of the gemcitabine pharmacokinetics. *BMC Syst. Biol.* 6:51. doi:10.1186/1752-0509-6-51

Lecca, P., and Priami, C. (2013). Biological network inference for drug discovery. *Drug Discov. Today* 18, 256–264. doi:10.1016/j.drudis.2012.11.001

Lee, D., Smallbone, K., Dunn, W. B., Murabito, E., Winder, C. L., Kell, D. B., et al. (2012). Improving metabolic flux predictions using absolute gene expression data. *BMC Syst. Biol.* 6:73. doi:10.1186/1752-0509-6-73

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

Lewis, N. E., Nagarajan, H., and Palsson, B. O. (2012). Constraining the metabolic genotype-phenotype relationship using a phylogeny of in silico methods. *Nat. Rev. Microbiol.* 10, 291–305. doi:10.1038/nrmicro2737

Link, H., Christodoulou, D., and Sauer, U. (2014). Advancing metabolic models with kinetic information. *Curr. Opin. Biotechnol.* 29, 8–14. doi:10.1016/j.copbio.2014.01.015

Liu, P.-K., and Wang, F.-S. (2008). Inference of biochemical network models in S-system using multiobjective optimization approach. *Bioinformatics* 24, 1085–1092. doi:10.1093/bioinformatics/btn075

Machado, D., and Herrgård, M. (2014). Systematic evaluation of methods for integration of transcriptomic data into constraint-based models of metabolism. *PLoS Comput. Biol.* 10:e1003989. doi:10.1371/journal.pcbi.1003989

Mahadevan, R., and Schilling, C. H. (2003). The effects of alternate optimal solutions in constraint-based genome-scale metabolic models. *Metab. Eng.* 5, 264–276. doi:10.1016/j.ymben.2003.09.002

Margolin, A. A., Nemenman, I., Basso, K., Wiggins, C., Stolovitzky, G., Dalla Favera, R., et al. (2006). ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC Bioinformatics* 7(Suppl. 1):S7. doi:10.1186/1471-2105-7-S1-S7

Millard, P., Massou, S., Wittmann, C., Portais, J.-C., and Létisse, F. (2014). Sampling of intracellular metabolites for stationary and non-stationary (13)C metabolic flux analysis in *Escherichia coli. Anal. Biochem.* 465C, 38–49. doi:10.1016/j.ab.2014.07.026

Mo, M. L., Palsson, B. O., and Herrgård, M. J. (2009). Connecting extracellular metabolomic measurements to intracellular flux states in yeast. *BMC Syst. Biol.* 3:37. doi:10.1186/1752-0509-3-37

Mourão, M. A., Srividhya, J., McSharry, P. E., Crampin, E. J., and Schnell, S. (2011). A graphical user interface for a method to infer kinetics and network architecture (MIKANA). *PLoS ONE* 6:e27534. doi:10.1371/journal.pone.0027534

Mueller, D., and Heinzle, E. (2013). Stable isotope-assisted metabolomics to detect metabolic flux changes in mammalian cell cultures. *Curr. Opin. Biotechnol.* 24, 54–59. doi:10.1016/j.copbio.2012.10.015

Müller, A., and Bockmayr, A. (2013). Fast thermodynamically constrained flux variability analysis. *Bioinformatics* 29, 903–909. doi:10.1093/bioinformatics/btt059

Nägele, T., Mair, A., Sun, X., Fragner, L., Teige, M., and Weckwerth, W. (2014). Solving the differential biochemical Jacobian from metabolomics covariance data. *PLoS ONE* 9:e92299. doi:10.1371/journal.pone.0092299

Navid, A., and Almaas, E. (2012). Genome-level transcription data of *Yersinia pestis* analyzed with a new metabolic constraint-based approach. *BMC Syst. Biol.* 6:150. doi:10.1186/1752-0509-6-150

Nemenman, I., Escola, G. S., Hlavacek, W. S., Unkefer, P. J., Unkefer, C. J., and Wall, M. E. (2007). Reconstruction of metabolic networks from high-throughput metabolite profiling data: in silico analysis of red blood cell metabolism. *Ann. N. Y. Acad. Sci.* 1115, 102–115. doi:10.1196/annals.1407.013

Nikerel, E., Berkhout, J., Hu, F., Teusink, B., Reinders, M. J. T., and de Ridder, D. (2012). Understanding regulation of metabolism through feasibility analysis. *PLoS ONE* 7:e39396. doi:10.1371/journal.pone.0039396

Oberhardt, M. A., Palsson, B. Ø., and Papin, J. A. (2009). Applications of genome-scale metabolic reconstructions. *Mol. Syst. Biol.* 5, 320. doi:10.1038/msb.2009.77

Öksüz, M., Sadıkoglu, H., and Çakır, T. (2013). Sparsity as cellular objective to infer directed metabolic networks from steady-state metabolome data: a theoretical analysis. *PLoS ONE* 8:e84505. doi:10.1371/journal.pone.0084505

Orth, J. D., Thiele, I., and Palsson, B. Ø (2010). What is flux balance analysis? *Nat. Biotechnol.* 28, 245–248. doi:10.1038/nbt.1614

Petranovic, D., and Nielsen, J. (2008). Can yeast systems biology contribute to the understanding of human disease? *Trends Biotechnol.* 26, 584–590. doi:10.1016/j.tibtech.2008.07.008

Postmus, J., Canelas, A. B., Bouwman, J., Bakker, B. M., van Gulik, W., de Mattos, M. J. T., et al. (2008). Quantitative analysis of the high temperature-induced glycolytic flux increase in *Saccharomyces cerevisiae* reveals dominant metabolic regulation. *J. Biol. Chem.* 283, 23524–23532. doi:10.1074/jbc.M802908200

Psychogios, N., Hau, D. D., Peng, J., Guo, A. C., Mandal, R., Bouatra, S., et al. (2011). The human serum metabolome. *PLoS ONE* 6:e16957. doi:10.1371/journal.pone.0016957

Quek, L.-E., Wittmann, C., Nielsen, L. K., and Krömer, J. O. (2009). OpenFLUX: efficient modelling software for 13C-based metabolic flux analysis. *Microb. Cell Fact.* 8, 25. doi:10.1186/1475-2859-8-25

Samoilov, M., Arkin, A., and Ross, J. (2001). On the deduction of chemical reaction pathways from measurements of time series of concentrations. *Chaos* 11, 108–114. doi:10.1063/1.1336499

Sauer, U. (2006). Metabolic networks in motion: 13C-based flux analysis. *Mol. Syst. Biol.* 2, 62. doi:10.1038/msb4100109

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

Schaub, J., Mauch, K., and Reuss, M. (2008). Metabolic flux analysis in *Escherichia coli* by integrating isotopic dynamic and isotopic stationary 13C labeling data. *Biotechnol. Bioeng.* 99, 1170–1185. doi:10.1002/bit.21675

Schmidt, H., Cho, K.-H., and Jacobsen, E. W. (2005). Identification of small scale biochemical networks based on general type system perturbations. *FEBS J.* 272, 2141–2151. doi:10.1111/j.1742-4658.2005.04605.x

Schmidt, M. D., Vallabhajosyula, R. R., Jenkins, J. W., Hood, J. E., Soni, A. S., Wikswo, J. P., et al. (2011). Automated refinement and inference of analytical models for metabolic networks. *Phys. Biol.* 8, 055011. doi:10.1088/1478-3975/8/5/055011

Schuetz, R., Kuepfer, L., and Sauer, U. (2007). Systematic evaluation of objective functions for predicting intracellular fluxes in *Escherichia coli. Mol. Syst. Biol.* 3, 119. doi:10.1038/msb4100162

Shlomi, T., Cabili, M. N., Herrgård, M. J., Palsson, B. Ø, and Ruppin, E. (2008). Network-based prediction of human tissue-specific metabolism. *Nat. Biotechnol.* 26, 1003–1010. doi:10.1038/nbt.1487

Smallbone, K., Simeonidis, E., Swainston, N., and Mendes, P. (2010). Towards a genome-scale kinetic model of cellular metabolism. *BMC Syst. Biol.* 4:6. doi:10.1186/1752-0509-4-6

Soh, K. C., and Hatzimanikatis, V. (2010). Network thermodynamics in the post-genomic era. *Curr. Opin. Microbiol.* 13, 350–357. doi:10.1016/j.mib.2010.03.001

Sorribas, A., Hernández-Bermejo, B., Vilaprinyo, E., and Alves, R. (2007). Cooperativity and saturation in biochemical networks: a saturable formalism using Taylor series approximations. *Biotechnol. Bioeng.* 97, 1259–1277. doi:10.1002/bit.21316

Srividhya, J., Crampin, E. J., McSharry, P. E., and Schnell, S. (2007). Reconstructing biochemical pathways from time course data. *Proteomics* 7, 828–838. doi:10.1002/pmic.200600428

Stanford, N. J., Lubitz, T., Smallbone, K., Klipp, E., Mendes, P., and Liebermeister, W. (2013). Systematic construction of kinetic models from genome-scale metabolic networks. *PLoS ONE* 8:e79195. doi:10.1371/journal.pone.0079195

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., Kurths, J., Fiehn, O., and Weckwerth, W. (2003). Observing and interpreting correlations in metabolomic networks. *Bioinformatics* 19, 1019–1026. doi:10.1093/bioinformatics/btg120

Sun, X., and Weckwerth, W. (2012). COVAIN: a toolbox for uni-and multivariate statistics, time-series and correlation network analysis and inverse estimation of the differential Jacobian from metabolomics covariance data. *Metabolomics* 8, 81–93. doi:10.1007/s11306-012-0399-3

Tarlak, F., Sadıkoglu, H., and Çakır, T. (2014). The role of flexibility and optimality in the prediction of intracellular fluxes of microbial central carbon metabolism. *Mol. Biosyst.* 10, 2459–2465. doi:10.1039/c4mb00117f

Teusink, B., Passarge, J., Reijenga, C. A., Esgalhado, E., van der Weijden, C. C., Schepper, M., et al. (2000). Can yeast glycolysis be understood in terms of in vitro kinetics of the constituent enzymes? Testing biochemistry. *Eur. J. Biochem.* 267, 5313–5329. doi:10.1046/j.1432-1327.2000.01527.x

Thiele, I., and Palsson, B. (2010). A protocol for generating a high-quality genome-scale metabolic reconstruction. *Nat. Protoc.* 5, 93–121. doi:10.1038/nprot.2009.203

Torralba, A. S., Yu, K., Shen, P., Oefner, P. J., and Ross, J. (2003). Experimental test of a method for determining causal connectivities of species in reactions. *Proc. Natl. Acad. Sci. U.S.A.* 100, 1494–1498. doi:10.1073/pnas.262790699

Toya, Y., Ishii, N., Hirasawa, T., Naba, M., Hirai, K., Sugawara, K., et al. (2007). Direct measurement of isotopomer of intracellular metabolites using capillary electrophoresis time-of-flight mass spectrometry for efficient metabolic flux analysis. *J. Chromatogr.* 1159, 134–141. doi:10.1016/j.chroma.2007.04.011

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

Van Kampen, N. G. (1992). *Stochastic Processes in Physics and Chemistry*. Amsterdam: Elsevier Science.

Van Winden, W. A., van Dam, J. C., Ras, C., Kleijn, R. J., Vinke, J. L., van Gulik, W. M., et al. (2005). Metabolic-flux analysis of *Saccharomyces cerevisiae* CEN.PK113-7D based on mass isotopomer measurements of (13)C-labeled primary metabolites. *FEMS Yeast Res.* 5, 559–568. doi:10.1016/j.femsyr.2004.10.007

Vance, W., Arkin, A., and Ross, J. (2002). Determination of causal connectivities of species in reaction networks. *Proc. Natl. Acad. Sci. U.S.A.* 99, 5816–5821. doi:10.1073/pnas.022049699

Varma, A., and Palsson, B. (1994). Stoichiometric flux balance models quantitatively predict growth and metabolic by-product secretion in wild-type *Escherichia coli* W3110: basic concepts, scientific and practical use. *Appl. Environ. Microbiol.* 60, 3724–3731.

Villaverde, A. F., Ross, J., Morán, F., and Banga, J. R. (2014). MIDER: network inference with mutual information distance and entropy reduction. *PLoS ONE* 9:e96732. doi:10.1371/journal.pone.0096732

Visser, D., Schmid, J. W., Mauch, K., Reuss, M., and Heijnen, J. J. (2004). Optimal re-design of primary metabolism in *Escherichia coli* using linlog kinetics. *Metab. Eng.* 6, 378–390. doi:10.1016/j.ymben.2004.07.001

Weckwerth, W., Loureiro, M. E., Wenzel, K., and Fiehn, O. (2004). Differential metabolic networks unravel the effects of silent plant phenotypes. *Proc. Natl. Acad. Sci. U.S.A.* 101, 7809–7814. doi:10.1073/pnas.0303415101

Weitzel, M., Nöh, K., Dalman, T., Niedenführ, S., Stute, B., and Wiechert, W. (2013). 13CFLUX2 – high-performance software suite for (13)C-metabolic flux analysis. *Bioinformatics* 29, 143–145. doi:10.1093/bioinformatics/bts646

Wiechert, W., Möllney, M., Petersen, S., and de Graaf, A. A. (2001). A universal framework for 13C metabolic flux analysis. *Metab. Eng.* 3, 265–283. doi:10.1006/mben.2001.0187

Wiechert, W., and Nöh, K. (2013). Isotopically non-stationary metabolic flux analysis: complex yet highly informative. *Curr. Opin. Biotechnol.* 24, 979–986. doi:10.1016/j.copbio.2013.03.024

Yizhak, K., Benyamini, T., Liebermeister, W., Ruppin, E., and Shlomi, T. (2010). Integrating quantitative proteomics and metabolomics with a genome-scale metabolic network model. *Bioinformatics* 26, i255–i260. doi:10.1093/bioinformatics/btq183

Young, J. D., Walther, J. L., Antoniewicz, M. R., Yoo, H., and Stephanopoulos, G. (2008). An elementary metabolite unit (EMU) based method of isotopically non-stationary flux analysis. *Biotechnol. Bioeng.* 99, 686–699. doi:10.1002/bit.21632

Zamboni, N., Fendt, S.-M., Rühl, M., and Sauer, U. (2009). (13)C-based metabolic flux analysis. *Nat. Protoc.* 4, 878–892. doi:10.1038/nprot.2009.58

Zamboni, N., Fischer, E., and Sauer, U. (2005). FiatFlux – a software for metabolic flux analysis from 13C-glucose experiments. *BMC Bioinformatics* 6:209. doi:10.1186/1471-2105-6-209

Zelezniak, A., Sheridan, S., and Patil, K. R. (2014). Contribution of network connectivity in determining the relationship between gene expression and metabolite concentration changes. *PLoS Comput. Biol.* 10:e1003572. doi:10.1371/journal.pcbi.1003572

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.

Received: 24 September 2014; accepted: 14 November 2014; published online: 03 December 2014.

Citation: Çakır T and Khatibipour MJ (2014) Metabolic network discovery by top-down and bottom-up approaches and paths for reconciliation. Front. Bioeng. Biotechnol. ***2****:62. doi: 10.3389/fbioe.2014.00062*

This article was submitted to Systems Biology, a section of the journal Frontiers in Bioengineering and Biotechnology.

frontiers in
BIOENGINEERING AND BIOTECHNOLOGY

ORIGINAL RESEARCH ARTICLE
published: 10 December 2014
doi: 10.3389/fbioe.2014.00068

Microalgal metabolic network model refinement through high-throughput functional metabolic profiling

Amphun Chaiboonchoe[1,2†], Bushra Saeed Dohai[1,2†], Hong Cai[1,2†], David R. Nelson[1,2], Kenan Jijakli[1,2,3] and Kourosh Salehi-Ashtiani[1,2*]

[1] Division of Science and Math, New York University Abu Dhabi, Abu Dhabi, UAE
[2] Center for Genomics and Systems Biology (CGSB), New York University Abu Dhabi Institute, Abu Dhabi, UAE
[3] Engineering Division, Biofinery, Manhattan, KS, USA

Edited by:
Kai Hua Zhuang, Technical University of Denmark, Denmark

Reviewed by:
Eduardo S. Zeron, Centro de Investigación y de Estudios Avanzados del Instituto Politécnico Nacional, Mexico
Eleftherios Pilalis, National Hellenic Research Foundation, Greece

***Correspondence:**
Kourosh Salehi-Ashtiani, Laboratory of Algal, Systems, and Synthetic Biology (LASSB), Division of Science and Math, Center for Genomics and Systems Biology (CGSB), P.O. Box 129188, New York University Abu Dhabi, Abu Dhabi, UAE
e-mail: ksa3@nyu.edu

†Amphun Chaiboonchoe, Bushra Saeed Dohai and Hong Cai have contributed equally to this work.

Metabolic modeling provides the means to define metabolic processes at a systems level; however, genome-scale metabolic models often remain incomplete in their description of metabolic networks and may include reactions that are experimentally unverified. This shortcoming is exacerbated in reconstructed models of newly isolated algal species, as there may be little to no biochemical evidence available for the metabolism of such isolates. The phenotype microarray (PM) technology (Biolog, Hayward, CA, USA) provides an efficient, high-throughput method to functionally define cellular metabolic activities in response to a large array of entry metabolites. The platform can experimentally verify many of the unverified reactions in a network model as well as identify missing or new reactions in the reconstructed metabolic model. The PM technology has been used for metabolic phenotyping of non-photosynthetic bacteria and fungi, but it has not been reported for the phenotyping of microalgae. Here, we introduce the use of PM assays in a systematic way to the study of microalgae, applying it specifically to the green microalgal model species *Chlamydomonas reinhardtii*. The results obtained in this study validate a number of existing annotated metabolic reactions and identify a number of novel and unexpected metabolites. The obtained information was used to expand and refine the existing COBRA-based *C. reinhardtii* metabolic network model iRC1080. Over 254 reactions were added to the network, and the effects of these additions on flux distribution within the network are described. The novel reactions include the support of metabolism by a number of D-amino acids, L-dipeptides, and L-tripeptides as nitrogen sources, as well as support of cellular respiration by cysteamine-*S*-phosphate as a phosphorus source. The protocol developed here can be used as a foundation to functionally profile other microalgae such as known microalgae mutants and novel isolates.

Keywords: microalgae, *Chlamydomonas reinhardtii*, flux balance analysis, phenotype microarray, metabolic network refinement

INTRODUCTION

Optimization of algal metabolism toward improved bioproduct production while maintaining strain robustness remains a challenge that requires experimental strategies informed through systems-level analyses of metabolism. The use of metabolic network models can guide the development of optimization strategies that would be otherwise difficult through rational designs (Oberhardt et al., 2009; Schmidt et al., 2010; Koskimaki et al., 2013; Koussa et al., 2014). While an increasing number of algal species are being isolated and sequenced for biofuel or other applications, to date, there are only a handful of reconstructed algal networks available (Koussa et al., 2014). A major obstacle in the reconstruction of high-quality network models for algae remains hinged on the inability to obtain rapid and high-throughput metabolic phenotypic data to guide and validate reconstruction efforts.

One potential high-throughput phenotypic analysis technology is the Biolog OmniLog® phenotype microarray (PM) (Biolog, Hayward, CA, USA) (Bochner et al., 2001; Bochner, 2003, 2009). By assaying cellular metabolism in response to thousands of metabolites, signaling molecules, and effector molecules (as well as osmolites), the Biolog PM assays have greatly boosted functional metabolic profiling by providing insight into function, metabolism, and environmental sensitivity (Bochner et al., 2001; Bochner, 2003, 2009). Biolog PM assays rely on the measurement of metabolite utilization of cells in 96-well microplates. Each well contains different nutrients, metabolites, and pH and osmolarity solutes. Other bioactive molecules such as antibiotics and hormones may also be assayed. This utilization is assessed and measured in the form of cell respiration determined by the amount of color development produced by the NADH reduction of a tetrazolium-based redox dye (Bochner et al., 2001; Bochner, 2003, 2009). Plates can be monitored automatically over time with the OmniLog platform. A common set of 20 96-well microplates are designed to measure carbon, nitrogen, sulfur, phosphorus utilization phenotypes, along with osmotic/ion, and pH effects. This high-throughput and standardized approach has the ability to provide a quick method for the phenotypic comparison of different strains and organisms in a convenient manner leading to insights into the metabolic state of the cell. While the PM technology has been used for metabolic phenotyping of various microbial species including bacteria and fungi, it has not been reported for the phenotyping of microalgae. Likewise, the technology has been successfully used for verification and expansion of a number of existing microbial metabolic network models (Bochner et al., 2001; Bochner, 2003, 2009; Bartell et al., 2014), yet its use for improvement of microalgal models remains unreported.

The goal of the present study is to establish a reliable method for characterizing metabolic phenotypes of microalgae that can be used to expand existing network models or guide the reconstruction of new algal metabolic models. We present the implementation of the PM platform for metabolic phenotyping of microalgae using *Chlamydomonas reinhardtii* as a model organism then expand a well-curated existing metabolic network model of *C. reinhardtii* accordingly.

MATERIALS AND METHODS

PHENOTYPE MICROARRAY EXPERIMENTS

Phenotyping was done using standard Biolog assay plates and using the OmniLog instrument. In total 190 substrate utilization assays for carbon sources (PM01 and PM02), 95 substrate utilization assays for nitrogen sources (PM03), 59 nutrient utilization assays for phosphorus sources, and 35 nutrient utilization assays for sulfur sources (PM04), along with peptide nitrogen sources (PM06–08) were utilized. A defined tris-acetate-phosphate (TAP) medium (Gorman and Levine, 1965) containing 0.1% tetrazolium violet dye "D" (Biolog, Hayward, CA, USA) was used for the PM tests. The carbon, nitrogen, phosphorus, or sulfur component of the media was omitted from the defined medium when applied to the respective PM microplates that tested for each of those sources.

Chlamydomonas reinhardtii strain CC-503 was obtained from the *Chlamydomonas* Resource Center at the University of Minnesota, USA. Cells were grown in fresh TAP media to mid-log phase, then spun down at 2,000 × *g* for 10 min, and then resuspended in fresh media to a final concentration of 1×10^6 cells before inoculation into Biolog's 96-well plates. A 100 μL aliquot of cell-containing media was inoculated into each well before the plates were inserted into the OmniLog system. A final concentration of 400 μL/mL timentin® (GlaxoSmithKline, New Zealand) was used to inhibit bacterial growth in all plates. In addition, ampicillin and kanamycin were used at 50–100 μg/ml occasionally. Bacterial contamination was monitored by streaking cells on yeast extract/peptone plates and performing gram stains before and after Biolog assays. All microplates were incubated at 30°C for up to 7 days and the dye color change (in the form of absorbance) was read with the OmniLog system every 15 min. As the OmniLog instrument does not provide a source of continuous light during incubation, the algae is assumed to be carrying out heterotrophic respiration.

DATA ANALYSIS

The Biolog PM data analysis was carried out using an OmniLog phenotype microarray (OPM) software package (Vaas et al., 2012, 2013) that runs within the R software environment. The raw kinetic data were exported as CSV files to the OPM package and then the biological information was added as metadata (e.g., strain designation, growth media, temperature, etc.). Kinetic curves were plotted from the raw data in the form of *xy* and level plots, and a statistical analysis was carried out to visualize the metabolic properties and generate OmniLog values. An OmniLog value or the curve parameter "A" simply lists the maximum height of the growth curve.

Duplicate assays were carried out for all the plates that were tested to assess reproducibility of the data. An assay was considered positive when the absorbance (OmniLog value) was positive after subtraction from the negative control well and the respective blank well. This summation is a representation of the abiotic reaction of the dye with the media in the presence of the tested compound.

IDENTIFICATION OF REACTIONS AND GENES ASSOCIATED WITH NEW METABOLITES

Gene to reaction associations for compounds were established as follows: assignment of a compound's enzyme commission number (EC) and relevant reactions were performed by searching KEGG[1] and MetaCyc[2]. The genomic evidence for each reaction was then recovered by using the identified EC numbers as a search basis in multiple available annotation resources from available algal annotation databases, such as the Joint Genome Institute (JGI), Phytozome[3], and peer-reviewed publications. When the query returned no genomic evidence for a given EC number, the relevant associated proteins in other organisms were identified then a profile-based search was carried out using the NCBI PSI-BLAST server with default settings and using non-redundant protein sequences (nr) in *C. reinhardtii* (taxid: 3055). PSI-BLAST hits with E values of ≤ 0.05 were manually curated for relevance to the searched EC number through either the evaluation of their described enzymatic activity, or by querying those BLAST hits through EMBL-EBI Pfam[4], or InterPro[5] protein domain prediction servers.

MODEL REFINEMENT AND EVALUATION

Identified reactions with their associated genes were added to iRC1080 using the COBRA Toolbox functions add Reaction and Change Gene Association. In addition, transport reactions for the new metabolites were incorporated into the model as transport by passive diffusion from the extracellular medium into the cytosol. The behavior of the new resultant model, iBD1106, was tested by carrying out flux balance analyses under light and dark conditions for the maximization of biomass as the objective function. The comparison of the two models was based on reported shadow prices (sensitivity of the objective function to changes in system variables) of the metabolites. The Biomass function was defined previously (Chang et al., 2011) for growth under dark and light conditions. The revised model can be found in the supplementary file iBD1106.xml in an SBML file format.

[1] http://www.genome.jp/kegg/
[2] http://metacyc.org/
[3] http://www.phytozome.net
[4] http://pfam.xfam.org/search
[5] http://www.ebi.ac.uk/interpro/

RESULTS

PHENOTYPE MICROARRAY SCREENING OF MODEL ALGA *CHLAMYDOMONAS REINHARDTII*

To implement the use of the PM platform for algal metabolic phenotyping, we used *C. reinhardtii* as a model. The single-cell green alga *C. reinhardtii* is a model organism that has been widely used for basic and applied biological research. Its genome was sequenced and publically released by JGI in 2007 (Merchant et al., 2007) and genome-scale models of its metabolism have been reconstructed (May et al., 2009; Chang et al., 2011; Dal'Molin et al., 2011). The ability to grow phototrophically or heterotrophically, along with rapid growth and scalability, are features that make this alga an attractive model system for algal-based biofuel studies.

Our pipeline (**Figure 1**) integrates the high-throughput PM assays, applied to the alga of interest, with genomic searches to provide experimental evidence that can lead to the refinement of an existing metabolic network model. The pipeline may also be applied for a new reconstruction if an existing model is not

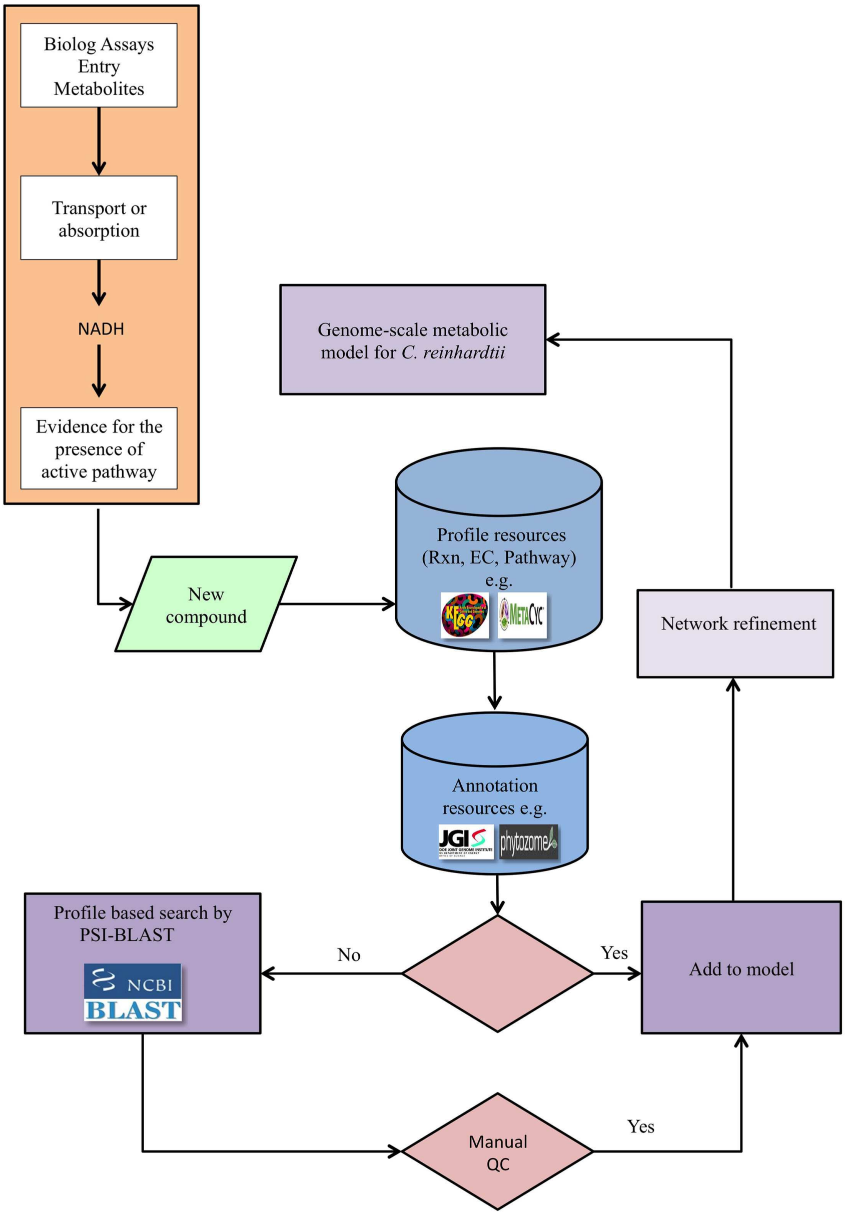

Biolog Assays Entry Metabolites
Transport or absorption
NADH
Evidence for the presence of active pathway
Genome-scale metabolic model for *C. reinhardtii*
New compound
Profile resources (Rxn, EC, Pathway) e.g.
KEGG
MetaCyc
Network refinement
Annotation resources e.g.
JGI
phytozome
Profile based search by PSI-BLAST
NCBI BLAST
No
Yes
Add to model
Manual QC
Yes

FIGURE 1 | The pipeline for genome-scale metabolic network refinement using PM data. After a new compound tests positive in a PM assay, its enzyme commission number (EC), reaction, and pathway are identified from available databases, e.g., KEGG and MetaCyc. Genomic evidence is then extracted directly from genomic and annotation resources when available and constitutes a link between genotype and phenotype. When direct genomic evidence is unavailable, the protein sequence is identified from the EC numbers and through the protein sequence, genetic evidence is identified via PSI-BLAST. The reconstructed metabolic network is then refined based on newly identified compounds, but only after a quality control step. The quality control step entails querying the protein domains using relevant databases.

available. The PM assays test the ability of the alga to utilize various carbon, nitrogen, sulfur, and phosphorus sources in a minimal medium. When a new compound tests positive for utilization, the compound's relevant reaction profiles are defined using metabolic knowledge bases such as KEGG (see text footnote 1) or MetaCyc (see text footnote 2). This step defines all potential reactions and pathways that can be associated with a metabolite to provide EC numbers. The next step is to find supporting genetic evidence from genetic databases specific to the alga, such as databases from the JGI, Phytozome (see text footnote 3), or peer-reviewed publications. If genetic evidence is available, the reactions and metabolites are added to the model to expand and refine the model. If, on the other hand, genomic evidence is not found in support of the EC number, a profile-based search, such as PSI-BLAST, can be performed to identify candidate genes associated with the reaction. The results of such searches are then manually evaluated; those passing this QC step are added to the network model. In exceptional cases, if genes are not identified for reactions but compelling biochemical evidence exists, reactions may be provisionally added to the network pending future investigations.

IMPLEMENTATION AND VALIDATION

We optimized the PM assays for metabolic profiling of *C. reinhardtii* by modifying the standard Biolog protocol with respect to inoculum concentration, type of dye, and pre-inoculation growth conditions (Materials and Methods). We used plates 1–4 and 6–8 of the PM platform, which provide a range of test compounds including utilization of carbon, nitrogen, sulfur, phosphorus, and a variety of di- and tripeptides. The summary kinetics of selected plates (PM01 and PM03) are shown in **Figure 2**. Splined-based curve fitting was implemented to extract the curve parameters [the lag phase (λ), the respiration (or growth rate μ or the steepness of the slope), the maximum cell respiration "A," and the area under the curve (AUC)]. The maximum cell respiration "A" of the blank and negative controls of each microwell plate (which represents abiotic reactivity of the dye with the medium and the test metabolite) were used as background subtraction values to identify positive metabolites. The "*xy*-plots" show the respiration measurements over time mapped to the assay 96-well plates, in terms of the raw measurements values (*y*-axis) and time (*x*-axis). In addition, the data was transformed to a heat map format to allow for a quick comparative overview of the multitude of the kinetic data. The heat map presents the kinetic values with different colors (varied from light yellow to dark orange or brownish; **Figures 2B,D**).

To assess the level of combined experimental and biological noise and systematic errors and biases from Biolog's PM measurements, the data from two independent replicate experiments were plotted against one another (**Figure 3**). This figure visually assesses the reproducibility of the PM data obtained from PM01–04 and PM10 plates. **Figure 3** shows that the majority of the data were identical as they fall on the 45° line with only a few outliers. This plot confirms the quality and reproducibility of the experiments for this alga.

IDENTIFICATION OF NEW METABOLITES

We compared the number of metabolites that can be identified by Biolog's PM (662 chemical compounds from seven plates {PM01–PM04, and PM06–PM08}) with the iRC1080 metabolites and the metabolites measured using gas chromatography time-of-flight (GC-TOF) (Bölling and Fiehn, 2005) (**Figure 4**). Only six metabolites were overlapping among the three sets (adenine, glycerol, glycine, myo-Inositol, putrescine, and uracil), while 149 were common between iRC1080 and the Biolog set under investigation. This shows that while each technology/tool has its strength in metabolic profiling research, the Biolog set can be a significant source of new metabolic information.

After subtracting the background signal, we observed acetic acid as the only positive assay for carbon utilization (in PM01 plate). Detection of acetate as the only carbon source from this plate is consistent with the *Chlamydomonas* literature (e.g., Harris, 2009) and provides evidence for specificity of our assays. Four positive reactions for sulfur utilization (sulfate, thiosulfate, tetrathionate, D,L-Lipoamide) and four positive assays for phosphorus utilization (thiophosphate, dithiophosphate, D-3-phospho-glyceric acid and cysteamine-*S*-phosphate) were detected. *C. reinhardtii* showed positive results for several nitrogen sources including both L-amino and D-amino acids, and less common amino acids such as L-homoserine, L-pyroglutamic acid, methylamine, ethylamine, ethanolamine, and D,L-α-amino-butyric acid. Furthermore, a large number of dipeptides and a few tripeptides assayed positive (**Table 1**).

Altogether, we identified 128 new metabolites from the PM data that were not present in our iRC1080 metabolic model: eight D-amino acids, tetrathionate, thiophosphate, dithiophosphate, cysteamine-*S*-phosphate, L-pyroglutamic acid, and ethylamine, 108 dipeptides, and 5 tripeptides. We note that sequence specificity was observed for utilization of both di- and tripeptides. The identified metabolites are summarized in **Table 1** and Table S2 in Supplementary Material.

We searched KEGG and MetaCyc to define all possible reactions and EC numbers associated with the identified new metabolites. Forty-nine unique EC numbers were associated with the newly identified metabolites. Table S2 in Supplementary Material includes pathways, reactions, EC numbers, proteins, and *Chlamydomonas* annotation sources for each of the metabolites. Five different sources were used to obtain genomic evidence for the reactions. These included Phytozome Version 10.0.2 (Goodstein et al., 2012), JGI Version 4 (Ghamsari et al., 2011), AUGUSTUS 5.0 and 5.2 (Chang et al., 2011), annotations from Manichaikul et al. (2009), and KEGG (Kanehisa et al., 2014). Out of 49 searched ECs, 15 transcripts could be found with annotations matching the searched ECs (**Table 1**; Tables S1 and S2 in Supplementary Material).

The metabolic reactions and their respective EC numbers for which no genomic evidence was found (using the aforementioned resources) were then entered into the Universal Protein Resource website (UniProt)[6] (Apweiler et al., 2004; Consortium, 2014). There, sequences that are related to the metabolites but are from other organisms were identified. Those sequences were then used to run Position-Specific Iterated BLAST (PSI-BLAST queries)[7]

[6] http://www.uniprot.org/

[7] https://blast.ncbi.nlm.nih.gov/Blast.cgi

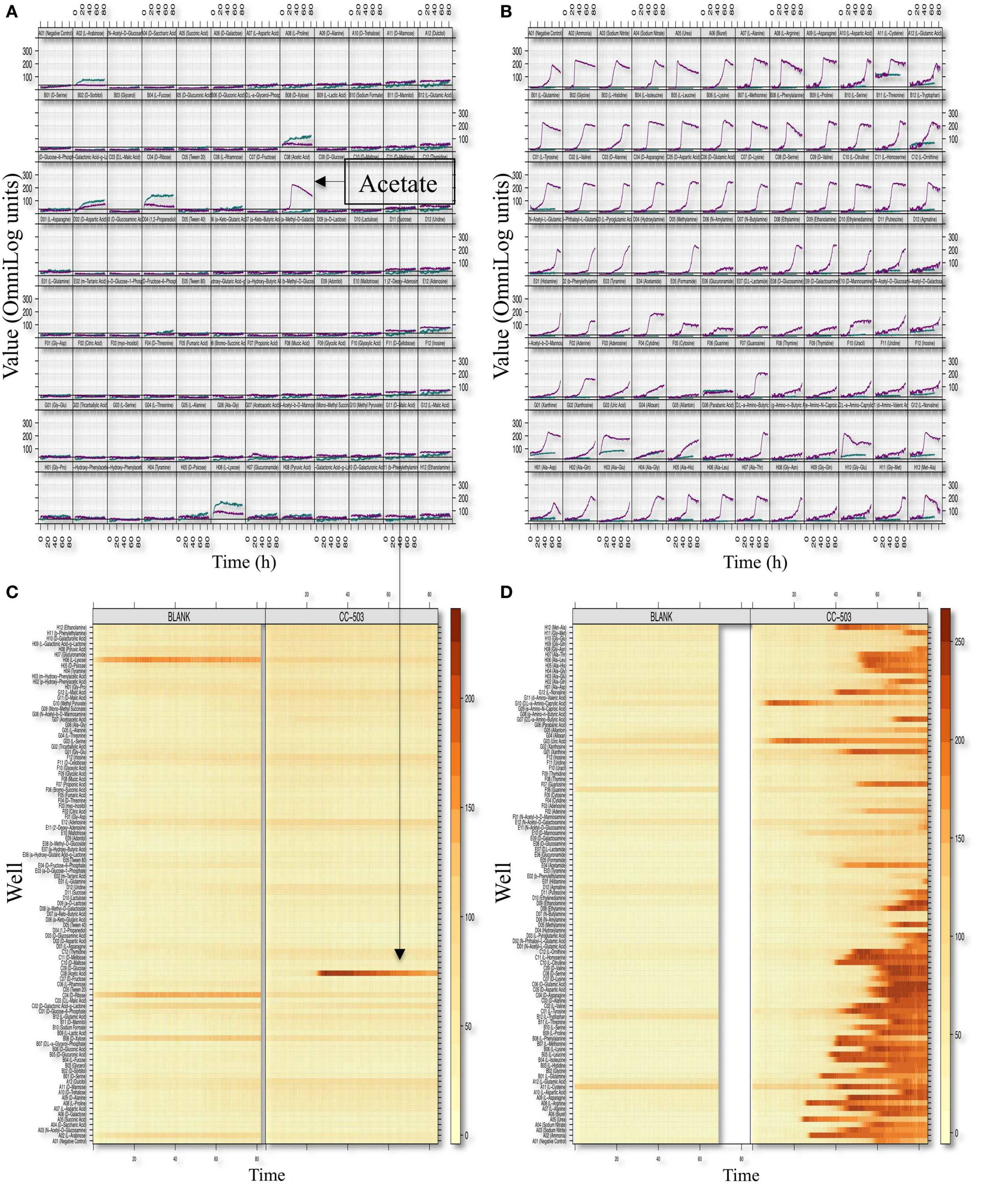


FIGURE 2 | Phenotypic microarray profiling selection of *C. reinhardtii*. Respiration (or growth) *xy*-plots and level plots of the PM01 [Carbon sources; **(A,B)**] and PM03 [Nitrogen sources; **(C,D)**] assay plates are shown. The figure is an 8 × 12 array where each cell represents a well plate and, thus, a given metabolite or growth environment. Within each cell or well representation, curves represent dye conversion by reduction (*y*-axis) as a function of time (*x*-axis). PM respiration curves from the CC-503 and blank are both shown in each cell and are indicated by color (teal color represents blank and purple color represents CC-503). The level-plot represents each respiration curve as a thin horizontal line changing color (or remaining unchanged) over time. Shading color changes from light yellow to dark orange or brownish based on the level of respiration measurement values, with the brownish color representing higher respiration values. Metabolites utilized by *C. reinhardtii* (CC-503) and the blank plates are shown.

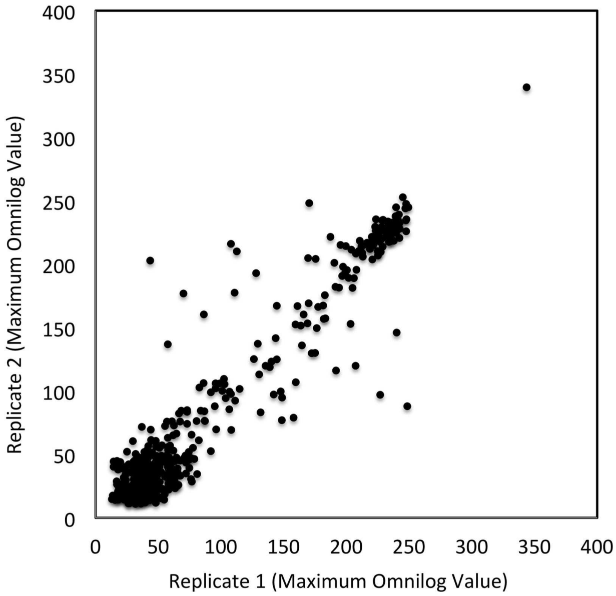


FIGURE 3 | Reproducibility of PM tests. OmniLog values were collected over a 168 h period and the maximum values were plotted for two replicate studies. Each axis represents the maximum OmniLog values for each study (the *x*-axis being one replicate study and the *y*-axis another). Identical values fall on a 45° line; there are a few deviating test values (some deviations were by more than 50 units). Each point represents a single maximum OmniLog value.

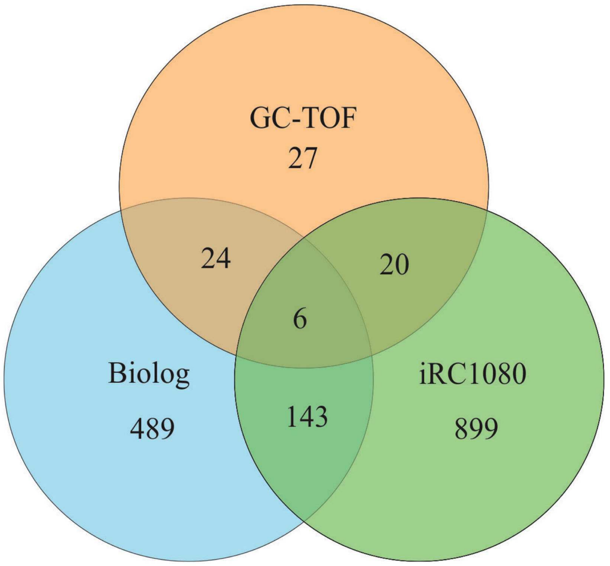


FIGURE 4 | Venn diagram of metabolites. The Venn diagram is a representation of metabolites common to Biolog's PM plates, the iRC1080 metabolic model and Gas Chromatography time-of-flight (GC-TOF) experiments. Each circle indicates the total number of metabolites that exists in each respective method of study, while the overlapping regions represent the number of metabolites shared between those methods of study. The iRC1080 metabolic model contains a total of 1,068 unique metabolites, the GC-TOF identified a total of 77 metabolites (Bölling and Fiehn, 2005), while there are a total of 662 metabolites tested using Biolog's PM plates.

Table 1 | List of identified positive substrate utilization metabolites (C, P, S, N) not present in the iRC1080 model.

Biolog chemical	EC[a]	Gene annotation	PSI-BLAST
Cysteamine-*S*-phosphate	3.1.3.1	JLM_162926[b,c,d,e]	
Tetrathionate	1.8.2.2		Insignificant E-value
	1.8.5.2		Insignificant E-value
D-Alanine	1.4.1.1		XP_001700222.1
	1.5.1.22		Failed manual QC
	2.1.2.7		Insignificant E-value
	1.4.3.3	Cre02.g096350.t1.3[f]	
	2.3.2.10		Insignificant E-value
	2.3.2.14		Insignificant E-value
	2.3.2.16		Insignificant E-value
	2.3.2.17		Insignificant E-value
	2.3.2.18		Insignificant E-value
	2.6.1.21		Failed manual QC
	3.4.13.22		XP_001698572.1, XP_001693532.1, XP_001701890.1, XP_001700930.1
	3.4.16.4	Chlre2_kg.scaffold_14000039[b,c,d]	
	3.4.17.8		Failed manual QC
	3.4.17.13		Insignificant E-value
	3.4.17.14		Insignificant E-value
	4.5.1.2		Insignificant E-value
	6.1.1.13		Failed manual QC
	6.1.2.1		Failed manual QC
	6.3.2.4	au.g14655_t1[b,c,d]	
	6.3.2.10		Failed manual QC
	6.3.2.16		Insignificant E-value
	6.3.2.35		Insignificant E-value
D-Asparagine	1.4.5.1		Insignificant E-value
	1.4.3.3	Cre02.g096350.t1.3[f]	
	3.1.1.96		Insignificant E-value
	2.3.1.36		Insignificant E-value
	1.4.99.1		XP_001692123.1
	3.5.1.77	e_gwW.1.243.1[b,c]	
	3.5.1.81		Insignificant E-value
	5.1.1.10		Failed manual QC
D-Aspartic acid	6.3.1.12		Insignificant E-value
	1.4.3.3	Cre02.g096350.t1.3[f]	
D-Glutamic acid	1.4.3.7		Insignificant E-value
	1.4.3.3		Insignificant E-value
D-Lysine	5.4.3.4		Insignificant E-value
	1.4.3.3	Cre02.g096350.t1.3[f]	
	6.3.2.37		Failed manual QC

(Continued)

Table 1 | Continued

Biolog chemical	EC[a]	Gene annotation	PSI-BLAST
D-Serine	2.7.11.8		Insignificant E-value
	2.7.11.17	Cre12.g486350.t1.3[b,c,d,e]	
	3.4.21.78		Failed manual QC
	3.4.21.104		Failed manual QC
	4.3.1.18	g6244.t1[e]	Failed manual QC
	6.3.2.35		Insignificant E-value
	6.3.3.5		Insignificant E-value
	1.4.3.3	Cre02.g096350.t1.3[f]	
D-Valine	1.21.3.1		Failed manual QC
	6.3.2.26		Failed manual QC
	1.4.3.3	Cre02.g096350.t1.3[f]	
L-Pyroglutamic acid			
Thiophosphate			
Dithiophosphate			
Ethylamine	6.3.1.6		
D,L-α-Amino-butyric acid	2.1.1.49		Insignificant E-value
	1.4.3.3	Cre02.g096350.t1.3[f]	
Di-peptide	3.4.13.18	Cre02.g078650.t1.3[b]	
Tri-peptide	3.4.11.4	Cre16.g675350.t1.3[b]	

[a] Reaction was not include if no gene was identified.

[b] Phytozome version 10.0.2 (http://phytozome.jgi.doe.gov/pz/portal.html#!info?alias=Org_Creinhardtii).

[c] JGI version 4 (Ghamsari et al., 2011).

[d] Augustus version 5 (Chang et al., 2011).

[e] KEGG (http://www.genome.jp/kegg/kegg1.html).

[f] JGI version 3.1 (Manichaikul et al., 2009).

Table 2 | Contents of iRC1080 and iBD1106.

Model	Reactions	Metabolites	Genes
iRC1080	2,191	1,706	1,086
iBD1106	2,445	1,959	1,106

Table 3 | Summery of new reactions in iBD1106.

Category or class of reactions	Number of reactions
Amino acids	20
Dipeptides	108
Tripeptides	5
Transport reaction	120

from the NCBI website to identify homologous sequences in *C. reinhardtii.* Only the sequences that produced significant alignments were considered; specifically, results with an E-value below 0.005 were retained. The final step before integrating the genes from the PSI-BLAST results with the iRC1080 metabolic model was to check whether the genes' relevant reactions related to the new metabolites; only hits with relevant annotated enzymatic reactions were kept. The PSI-BLAST yielded four additional transcripts (**Table 1**; Table S2 in Supplementary Material).

MODEL REFINEMENT

The metabolites identified as new to the network were categorized and annotated in the model based on their utilization into nitrogen sources, phosphate sources, and sulfur sources. The nitrogen source metabolites were 8 D-amino acids, 2 L-amino acids, 108 dipeptides, and 5 tripeptides. The phosphate sources were cysteamine-*S*-phosphate, thiophosphate, and dithiophosphate. The only new sulfur source metabolite was tetrathionate. No genomic evidence for tetrathionate was found in databases and its PSI-BLAST E values did not pass the threshold of 0.005, thus, no reaction for this metabolite was added to the model. In addition, L-pyroglutamic acid, thiophosphate, dithiophosphate, and ethylamine were not added to model due to lack of genomic evidence.

To expand the existing model, reactions associated with the new metabolites and the genes associated with the new reactions were added to iRC1080 model to generate an expanded network, iBD1106. iBD1106 accounts for 2,445 reactions, 1,959 metabolites, and 1,106 genes (**Table 2**). The new 254 added reactions are distributed as follows: 20 amino acid reactions, 108 di-peptide reactions, 5 tri-peptide reactions, and 120 transport reactions (**Table 3**). The new 20 amino acids reactions were associated with 4 new genes (Cre02.g096350.t1.3, au.g14655_t1, e_gwW.1.243.1, Cre12.g486350.t1.3). The D-amino acids are oxidized into ammonium and a 2-oxo-carboxylate via the following reaction with EC number of 1.4.3.3 and associated gene Cre02.g096350.t1.3:

$$\mathrm{D-amino\ acid} + \mathrm{O_2} + \mathrm{H_2O} \rightarrow \mathrm{NH_4} + \mathrm{H_2O_2} + 2 - \mathrm{oxo\ carboxylate} \tag{1}$$

Equation 1 is a general reaction for all D-amino acids. However, some D-amino acids contribute to different reactions in addition to their own oxidation reactions. For example, D-serine reacts with ATP producing ADP and phospho-D-serine. Moreover, the chirality of D-amino acids can also be inverted into L forms and vise versa through annotated racemases (Table S2 in Supplementary Material).

Four genes identified by PSI-BLAST were added into the model and account for the D-alanine transaminase reaction (Eq. 2); XP_001698572.1, XP_001693532.1, XP_001701890.1, XP_001700930.1:

$$2 - \mathrm{oxoglutarate} + \mathrm{D} - \mathrm{alanine} \leftrightarrow \mathrm{D} - \mathrm{glutamate} + \mathrm{pyruvate} \tag{2}$$

In addition, XP_001692123.1, a PSI-BLAST identified gene, was associated with the oxidation of D-asparagine reaction as shown in Eq. 1.

A total of 113 added new reactions account for the hydrolysis of dipeptides and tripeptides. The hydrolysis of dipeptides and tripeptides are associated with two genes, one for dipeptides (Cre02.g078650.t1.3), and one for tripeptides (Cre16.g675350.t1.3). The dipeptides and tripeptides are decomposed into their unit L-amino acids, for instance, Leu–Pro decomposes into L-leucine and L-proline.

With respect to sources of phosphorus, a reaction for hydrolysis of cysteamine-*S*-phosphate into cysteamine and phosphate was added according to the following reaction that is associated with the gene JLM_162926:

$$\text{Cysteamine} - \text{S} - \text{Phosphate} + \text{H}_2\text{O} \rightarrow \text{Cysteamine} + \text{Phosphate} \quad (3)$$

In order to specify the cellular compartment where the new reactions occur, we used the WoLF PSORT tool (Horton et al., 2007)[8] and the results reported by Ghamsari et al. (2011). By providing protein sequences that are associated with the new reactions, WoLF PSORT predicted that the new reactions are localized to the cytosol.

In metabolic models, incomplete biochemical information may create gaps that form discontinuity in the network. In order to identify if any new gaps were introduced in the new model, gapFind, a COBRA command that lists root gaps, was used. The root gaps are defined as metabolites that cannot be produced in the metabolic model (Becker et al., 2007; Schellenberger et al., 2011). Using this command we found that both iRC1080 and iBD1106 models contain the same 91 root gaps. This indicates that the addition of the new metabolites and their associated reactions, did not introduce any new gaps. We note that transport reactions for the import of new metabolites into the cytosol were added.

The metabolic behavior of iBD1106 was tested under light conditions (no acetate) and dark conditions (with acetate) by carrying out flux balance analyses with the biomass as the maximized objective function. To assess the contribution that each metabolite makes to the set objective function, shadow prices for all metabolites were obtained (Tables S3 and S4 in Supplementary Material). The shadow price of a metabolite is defined as the change in an objective function with respect to flux changes of a metabolite (Varma et al., 1993; Orth et al., 2010). Shadow price allows the determination of whether a metabolite is in "excess" or is "limiting" the objective function, e.g., biomass production. Negative values are for metabolites that will decrease the objective function, positive values are for those that will increase the objective function, and values of 0 are for metabolites that will have no effect on the objective function. The comparison of shadow prices between iBD1106 and iRC1080 indicates that, for most metabolites, a large change is not observed (**Figure 5**; Tables S3–S5 in Supplementary Material); however, differences are observed in 105 and 70 cases under light and dark growth, respectively. Instances of such metabolites are provided in **Table 4**.

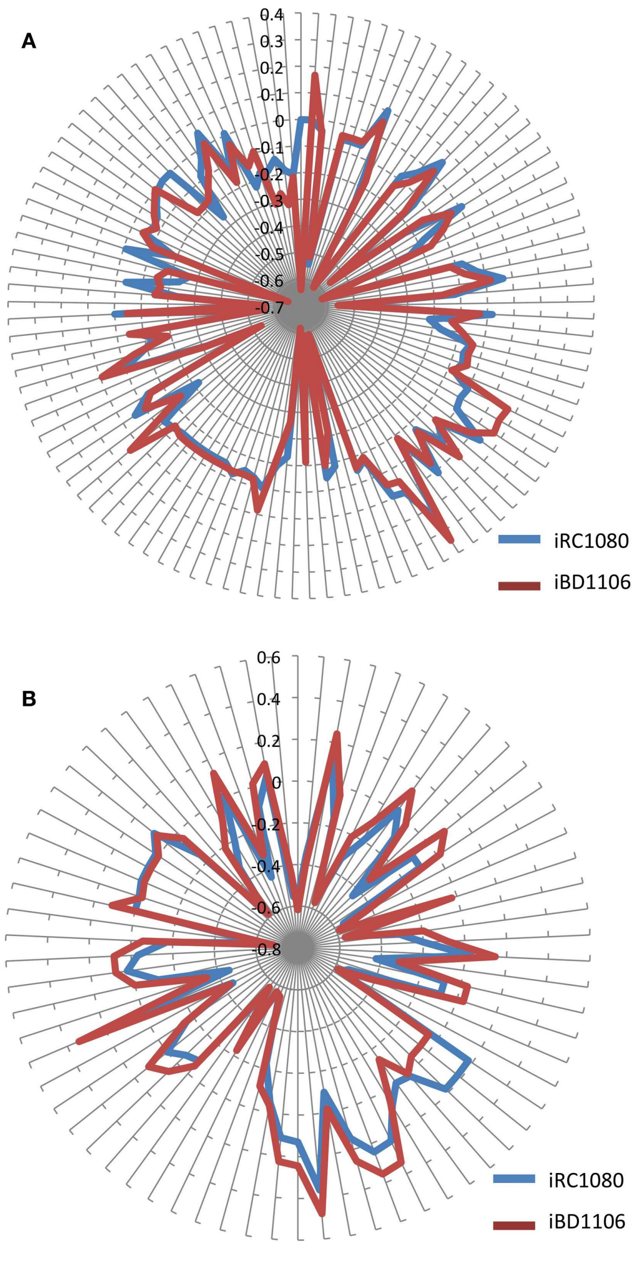


FIGURE 5 | Shadow prices of metabolites in iRC1080 and iBD1106 under different conditions for biomass maximization. Each circle on the "radar plots" corresponds to a shadow price value, while each line extending from the center of a plot indicates a metabolite. **(A)** shows the different shadow prices and subsequently metabolic behaviors of iRC1080 and iBD1106 under light growth conditions, while **(B)** shows the different metabolic behaviors of metabolites of iRC1080 and iBD1106 under dark growth conditions.

DISCUSSION

Algae are a group of diverse photosynthetic eukaryotes, which are polyphyletic in origin (Pröschold and Leliaert, 2007). Algal lineages include the viridiplantae, which the green algae (or Chlorophyta) belong to; stramenopiles that include brown, golden, and yellow algae and diatoms; rhodophyta or the red algae; and photosynthetic alveolates that include dinoflagelates (Barton et al., 2007). Given the evolutionary distances between these lineages, significant differences in genome size and coding potential,

[8] http://www.genscript.com/psort/wolf_psort.html

Table 4 | Examples of significant shadow prices for iRC1080 and iBD1106.

Growth condition	Metabolite	Name	iRC1080	iBD1106
Light	4r5au	4-(1-d-Ribitylamino)-5-aminouracil	0	0.168
	5aprbu	5-Amino-6-(5′-phosphoribitylamino)uracil	−0.009	0.158
	pa1819Z18111Z	1-(9Z)-octadecenoyl,2-(11Z)-octadecenoyl-sn-glycerol3-phosphate	−0.009	−0.65
Dark	4abut	4-aminobutanoate	0.18	−0.05

environmental niche, and metabolic properties can be expected. Members of green algae may be aquatic or soil organisms with mixotrophic or autotrophic modes of metabolism (Pröschold and Leliaert, 2007). In addition, microalgae may or may not require co-factors for their growth. Studies on microalgal growth requirements have indicated that more than half require cobalamin (vitamine B12), while 22% require thiamine and 5% need biotin (Croft et al., 2006). Interestingly, these requirements are not reflected in algal phylogeny (Helliwell et al., 2011).

Genomic approaches powered by next-generation sequencing technologies help to improve the understanding of the encoded algal metabolic potential; however, the full characterization of algal metabolism requires phenotypic data. For instance, the metabolome of *C. reinhardtii* has been studied under a number of conditions, including sulfur deprivation (Matthew et al., 2009; Shu and Hu, 2012; Aksoy et al., 2013), nitrogen deprivation (Blaby et al., 2013; Courant et al., 2013), and response to irradiance (Mettler et al., 2014) to provide insight into regulatory and metabolic responses of the species to environmental perturbations. In addition, transcriptomics, proteomics, and metabolomics studies have guided non-targeted profiling approaches for the detection and quantification of metabolites. Those non-targeted profiling approaches have included nuclear magnetic resonance (NMR), liquid chromatography mass spectrometry (LC-MS/MS), and gas chromatography mass spectrometry (GC/MS) (Veyel et al., 2014; Wase et al., 2014). The ability to study functional responses and phenotypes has been classically limited to targeted serial studies that usually employ mutagenesis, genetic knockouts, genetic over-expression, and physiological studies (Bochner, 2003; Dent et al., 2005; Morgan et al., 2009; Tshikhudo et al., 2013; Greetham, 2014). The wealth of phenotypic information gained from the PM technology, as demonstrated in this study, can help provide more complete systems-level knowledge when combined with other omics data, and help develop and refine metabolic models.

Genome-scale metabolic networks provide predicted genotype-phenotype relationships through metabolic flux optimization-based approaches. We previously reconstructed a genome-scale model for *C. reinhardtii* (iRC1080) (Chang et al., 2011) based on literature evidence (entailing ~250 sources), structurally verified genomic evidence, and predicted gene function and cellular localization information. This model has 1,706 metabolites with 2,191 reactions. Through the pipeline that we have described in this work, we were able to expand the network significantly to include 1,959 metabolites, 2,445 reactions, and 1,106 associated genes. A clear advantage that the PM provides is functional assays for entry metabolites to inform model refinement. Whereas mass spectrometry approaches give information on intermediate- and final-level metabolites, PM assays have the unique capability, due to the accounting for entry-level metabolites, to inform more complete models from the start of metabolic pathways. PM assays and mass spectrometry can therefore be considered as complementary approaches when characterizing organisms' metabolic profiles, with each technology refining and filling in specific gaps in metabolic models. Yet, PM's contribution to a metabolic model's refinement is made through a rapid, high-throughput, and convenient manner with an entire set of metabolites assayed in 5–7 days.

NEW METABOLITES

We have identified a number of di and tripeptides, and D-amino acids that significantly expand the list of nitrogen utilization compounds in *C. reinhardtii*. While we found D-amino acids can support metabolism of *C. reinhardtii*, they may be involved in additional functions. A serine-type D-alanyl-D-alanine carboxypeptidase was found in *C. reinhardtii*'s genome that could potentially be involved in D-alanine metabolism. Serine-type D-alalyl-D-alanine carboxypeptidases have been shown to play a variety of protective roles including protection against ionic and hyperosmotic stress (Príncipe et al., 2009). A D-alanine ligase was found in *C. reinhardtii*'s genome that is potentially involved in D-alanine multimerization. Recent research using ^{15}N NMR spectroscopy found that D-alanine accumulated in plants during UV exposure and this finding is supported by previous research under various stress signals (Monselise et al., 2014). Therefore, the possibility that D-amino acids might have additional cellular functions in *C. reinhardtii*, aside from providing a source of nitrogen, can be a subject of future investigations.

Chlamydomonas reinhardtii is known to be able to use a variety of amino acids as a sole nitrogen source as long as acetate is present (Munoz-Blanco et al., 1990). In *C. reinhardtii*, arginine is the only amino acid known to be imported with high affinity; the rest are believed to be deaminated extracellularly (Kirk and Kirk, 1978) or transported passively (Zuo et al., 2012). We note that a search in the literature for D-amino acid transports has not provided any information on the mode of transport for this class of amino acids in *C. reinhardtii*, nor is it known if the *C. reinhardtii* deaminase can deaminate D-amino acids. However, *C. reinhardtii* has been shown to exhibit amino acid racemerase activity (Takahashi et al., 2012), which could explain the ability to assimilate D-amino acids intracellularly. This also provides indirect evidence that these amino acids may be absorbed or transported into the cell for conversion to their L counterparts. A biological function for D-amino acids has not been clearly defined; however, D-alanine and D-aspartate were detected in algae using a reversed-phase HPLC; D-alanine was present in some marine diatoms while D-aspartate was found in all the selected freshwater green microalgae and marine diatoms (Yokoyama et al., 2003).

In many microbes, dipeptides are imported into the intracellular compartment before they are eventually hydrolyzed. For instance, *Francisella tularensis* relies on an amino acid transporter of the major facilitator superfamily of secondary transporters for transporting amino acids intracellularly. Furthermore, dipeptides containing asparagine were effective at restoring cellular multiplication in the infection cycle of a *F. tularensis* mutant that lacked that essential amino acid transporter (Gesbert et al., 2014). In this study, a variety of dipeptides were found to promote heterotrophic respiration in *C. reinhardtii*. The latest version of *C. reinhardtii*'s genome contains a gene annotated as a peptide hydrolase Cre02.g078650.t1.3. We note that the detected utilization of the dipeptides is not without sequence specificity as 159 out of 267 of the dipeptides and 9 out of 14 of the tripeptides did not result in positive assay results.

From these newly identified metabolites, three phosphorus compounds were found: (1) cysteamine-*S*-phosphate ($C_2H_7NO_3PS$), which is an organic phosphorothioate anion that is derived from deprotonation of thiophosphate OH groups and protonation of the amino group, (2) thiophosphate (or phosphorothioate), and (3) dithiophosphate, which is the product of the reaction of a base with phosphorus pentasulfide.

The only new sulfur source that was identified, tetrathionate, is a sulfur oxoanion and is derived from the compound tetrathionic acid and is commonly found in soils. It is a key intermediate in the oxidation of various reduced inorganic sulfur compounds. Several species of bacteria including *Salmonella enterica* (Winter et al., 2010) and *Acidithiobacillus ferrooxidans* (Rohwerder et al., 2003; Holmes and Bonnefoy, 2007; Chen et al., 2012) are known to be able to assimilate tetrathionate. Strains of *A. ferrooxidans* overexpressing tetrathionate hydrolase (tetH) were found to grow better on both sulfur and tetrathionate. In the archeon *Acidianus hospitalis*, tetrathionate is secreted to form filaments from tetrathionate homomultimers (Krupovic et al., 2012). These remarkable filaments are believed to play a role in sulfur metabolism and adaptation to *A. hospitalis*'s extreme environment. Prokaryotes have also been shown to use tetrathionate as an electron acceptor in cobalamin (coenzyme B12) synthesis (Roth et al., 1996). Sulfur is commonly assimilated as reduced sulfur for most living organisms, but bacteria are known to reduce tetrathionate, thiosulfate, sulfite, sulfur, and dimethyl sulfoxide in dissimilatory reactions as well (Barrett and Clark, 1987). Tetrathionate is often used as an electron sink for oxidative phosphorylation (Chen et al., 2012). Bacteria that are known to respire using tetrathionate are often found to have the capability of reducing thiosulfate as well, but thiosulfate is not found to be reduced among organisms that do not respire thiosulfate (Rohwerder et al., 2003). Considering that *C. reinhardtii* is a soil organism, the ability to assimilate this compound is likely to provide an important survival advantage in *Chlamydomonas*' natural life cycle.

iBD1106 MODEL VS. iRC1080

Different behaviors can be observed for iBD1106 than those for iRC1080 under different conditions. When the biomass production was set as the objective function, a differential change can be noticed as a result of growth conditions. The addition of the new nitrogen sources (D-amino acids, dipeptides, and tripeptides) has a significant and differential effect on the shadow prices of metabolites under light and dark conditions for biomass production (**Figures 5A,B**, respectively).

Under light growth, the D-aspartate in iBD1106 showed significant effect on the behavior of the chloroplastic metabolites of the riboflavin pathway. In iBD1106, D-aspartate is converted into L-aspartate through racemase, and L-aspartate can be produced through hydrolysis of its dipeptides (Asp–Leu, Asp–phe, Pro–Asp, Asp–Ala, Asp–Gln, and Asp–Gly). Also the oxidation of D-asparagine produces D-aspartate as oxocarboxylate (Eq. 1). The addition of L-aspartate increases its consumption in purine metabolism, which yields to more production of 2,5-Diamino-6-hydroxy-4-(5′-phosphoribos ylamino)-pyrimidine (25dhpp). The latter can be converted into 5-Amino-6-(5′-phosphoribosylamino)uracil (5apru) in the riboflavin metabolism resulting in an excess of 4-(1-D-Ribitylamino)-5-aminouracil (4r5au) and 5aprbu, with shadow prices of 0.168 and 0.158, respectively. Those results were not observed in iRC1080.

Another example of model discrepancy under light growth is the effect of adding D-serine reactions in iBD1106. Addition of D-serine limited the availability of the metabolite 1-(9Z)-octadecenoyl,2-(11Z)-octadecenoyl-sn-glycerol-3-phosphate (pa1819Z18111Z) (shadow price −0.009 in iRC1080 and −0.65 in iBD1106). This metabolite is produced and consumed by the reactions of glycerolipid metabolism for the production of Palmitoyl-CoA (n-C16:0CoA) (pmtcoa). The addition of L-serine in iBD1106 results in more consumption of pmtcoa in the sphingolipid metabolism through the reaction serine C-palmitoyltransferase (SERPT) that produces 3-dehydrosphinganine.

Under dark growth conditions, 4-aminobutanoate was in excess in iRC1080 and became limiting in iBD1106 with shadow price values of 0.18 and −0.05, respectively. The reason for this limiting availability is the addition of D-histidine and D-glutamate dipeptides hydrolysis reactions, e.g., Ala–His, and inversion into L-histidine and L-glutamate through a racemase. This addition increases the consumption of L-glutamate and L-histidine along with 4-aminobutanoate in glutamate and arginine and proline metabolisms, respectively. Moreover, the dark growth condition did not affect the behavior of 4-aminobutanoate significantly in iBD1106; however, in iRC1080 it was shifted from a limiting metabolite (−0.07) into an excess metabolite (0.18) (**Table 4**). The excessiveness of 4-aminobutanoate in iRC1080 under dark conditions might be related to the high consumption of NADPH under dark growth conditions. In proline metabolism, NADPH and 4-aminobutanoate are consumed more rapidly in dark than that in light conditions. As such, the addition of D-histidine and D-glutamate compensates the effect of growth under dark in the proline metabolism.

CONCLUSION

Phenotypic profiling has tremendous utility in modeling and understanding algal metabolism and is essential in elucidating genotypic differences in algae and the effects of environmental conditions on metabolism. The method presented here demonstrates the first reproducible study utilizing PM assays in profiling microalgae using *C. reinhardtii* as a model. We observed positive growth on 148 nutrients (one positive assay for C-source utilization, four positive assays for the S-source and P-source utilization, and 139 positive assays for N-source utilization). The wealth of phenotypic data can be used along with other references to compare organisms with known mutants or unknown isolates. This wealth of information will also shed light on new and novel metabolic pathways. The substrate utilization information and the newly identified metabolites were used for metabolic network expansion and refinement of the iRC1080 metabolic model. The study also provides a framework to bridge the missing links between genomics and metabolomics in microalgae. The described work provides an excellent method for the initial characterization of newly isolated or uncharacterized strains of algae. This combination of high-throughput phenotypic screening with metabolic modeling can allow for rapid refinement of existing metabolic network models as demonstrated and also provides biochemical evidence to support *de novo* reconstruction of new algal models.

ACKNOWLEDGMENTS

Major support for this work was provided by New York University Abu Dhabi Institute grant G1205-1205i and NYU Abu Dhabi Faculty Research Funds (AD060). We thank Basel Khraiwesh for help in designing some of the figures.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at http://www.frontiersin.org/Journal/10.3389/fbioe.2014.00068/abstract

Data Sheet S1:

Table S1 | PM assays for PM01 to PM04 and PM06 to PM08, positive assays are shaded with yellow.

Table S2 | A table of new metabolites that were added to generate iBD1106 along with their genetic annotations.

Table S3 | Shadow prices for metabolites of iRC1080 and iBD1106 when biomass was set as objective function under growth with light and no acetate.

Table S4 | Shadow prices for metabolites of iRC1080 and iBD1106 when biomass was set as objective function with growth under dark with acetate.

Table S5 | Shadow prices for new metabolites in iBD1106 when Biomass was set as the objective function (growth under light without acetate and under dark with acetate).

Data Sheet S2 | *Chlamydomonas reinhardtii* metabolic network model, iBD1106, in SBML format.

Figure S1 | Phenotype microarray results for plates 1–4, and 6–8.

REFERENCES

Aksoy, M., Pootakham, W., Pollock, S. V., Moseley, J. L., González-Ballester, D., and Grossman, A. R. (2013). Tiered regulation of sulfur deprivation responses in *Chlamydomonas reinhardtii* and identification of an associated regulatory factor. *Plant Physiol.* 162, 195–211. doi:10.1104/pp.113.214593

Apweiler, R., Bairoch, A., Wu, C. H., Barker, W. C., Boeckmann, B., Ferro, S., et al. (2004). UniProt: the universal protein knowledgebase. *Nucleic Acids Res.* 32, D115–D119. doi:10.1093/nar/gkh131

Barrett, E. L., and Clark, M. A. (1987). Tetrathionate reduction and production of hydrogen sulfide from thiosulfate. *Microbiol. Rev.* 51, 192.

Bartell, J. A., Yen, P., Varga, J. J., Goldberg, J. B., and Papin, J. A. (2014). Comparative metabolic systems analysis of pathogenic *Burkholderia*. *J. Bacteriol.* 196, 210–226. doi:10.1128/JB.00997-13

Barton, N. H., Briggs, D. E. G., Eisen, J. A., Goldstein, D. B., and Patel, N. H. (2007). *Evolution*. Cold Spring Harbor, NY: Cold Spring Harbor Laboratory Press.

Becker, S. A., Feist, A. M., Mo, M. L., Hannum, G., Palsson, B. Ø, and Herrgard, M. J. (2007). Quantitative prediction of cellular metabolism with constraint-based models: the COBRA Toolbox. *Nat. Protoc.* 2, 727–738. doi:10.1038/nprot.2007.99

Blaby, I. K., Glaesener, A. G., Mettler, T., Fitz-Gibbon, S. T., Gallaher, S. D., Liu, B., et al. (2013). Systems-level analysis of nitrogen starvation-induced modifications of carbon metabolism in a *Chlamydomonas reinhardtii* starchless mutant. *Plant Cell* 25, 4305–4323. doi:10.1105/tpc.113.117580

Bochner, B. R. (2003). New technologies to assess genotype-phenotype relationships. *Nat. Rev. Genet.* 4, 309–314. doi:10.1038/nrg1046

Bochner, B. R. (2009). Global phenotypic characterization of bacteria. *FEMS Microbiol. Rev.* 33, 191–205. doi:10.1111/j.1574-6976.2008.00149.x

Bochner, B. R., Gadzinski, P., and Panomitros, E. (2001). Phenotype microarrays for high-throughput phenotypic testing and assay of gene function. *Genome Res.* 11, 1246–1255. doi:10.1101/gr.186501

Bölling, C., and Fiehn, O. (2005). Metabolite profiling of *Chlamydomonas reinhardtii* under nutrient deprivation. *Plant Physiol.* 139, 1995–2005. doi:10.1104/pp.105.071589

Chang, R. L., Ghamsari, L., Manichaikul, A., Hom, E. F., Balaji, S., Fu, W., et al. (2011). Metabolic network reconstruction of *Chlamydomonas* offers insight into light-driven algal metabolism. *Mol. Syst. Biol.* 7, 518. doi:10.1038/msb.2011.52

Chen, L., Ren, Y., Lin, J., Liu, X., Pang, X., and Lin, J. (2012). *Acidithiobacillus caldus* sulfur oxidation model based on transcriptome analysis between the wild type and sulfur oxygenase reductase defective mutant. *PLoS ONE* 7:e39470. doi:10.1371/journal.pone.0039470

Consortium, T. U. (2014). Activities at the universal protein resource (UniProt). *Nucleic Acids Res.* 42, D191–D198. doi:10.1093/nar/gkt1140

Courant, F., Martzolff, A., Rabin, G., Antignac, J.-P., Le Bizec, B., Giraudeau, P., et al. (2013). How metabolomics can contribute to bio-processes: a proof of concept study for biomarkers discovery in the context of nitrogen-starved microalgae grown in photobioreactors. *Metabolomics* 9, 1286–1300. doi:10.1007/s11306-013-0532-y

Croft, M. T., Warren, M. J., and Smith, A. G. (2006). Algae need their vitamins. *Eukaryot. Cell* 5, 1175–1183. doi:10.1128/EC.00097-06

Dal'Molin, C. G., Quek, L.-E., Palfreyman, R., and Nielsen, L. (2011). AlgaGEM – a genome-scale metabolic reconstruction of algae based on the *Chlamydomonas reinhardtii* genome. *BMC Genomics* 12:S5. doi:10.1186/1471-2164-12-S4-S5

Dent, R. M., Haglund, C. M., Chin, B. L., Kobayashi, M. C., and Niyogi, K. K. (2005). Functional genomics of eukaryotic photosynthesis using insertional mutagenesis of *Chlamydomonas reinhardtii*. *Plant Physiol.* 137, 545–556. doi:10.1104/pp.104.055244

Gesbert, G., Ramond, E., Rigard, M., Frapy, E., Dupuis, M., Dubail, I., et al. (2014). Asparagine assimilation is critical for intracellular replication and dissemination of *Francisella*. *Cell. Microbiol.* 16, 434–449. doi:10.1111/cmi.12227

Ghamsari, L., Balaji, S., Shen, Y., Yang, X., Balcha, D., Fan, C., et al. (2011). Genome-wide functional annotation and structural verification of metabolic ORFeome of *Chlamydomonas reinhardtii*. *BMC Genomics* 12:S4. doi:10.1186/1471-2164-12-S1-S4

Goodstein, D. M., Shu, S., Howson, R., Neupane, R., Hayes, R. D., Fazo, J., et al. (2012). Phytozome: a comparative platform for green plant genomics. *Nucleic Acids Res.* 40, D1178–D1186. doi:10.1093/nar/gkr944

Gorman, D. S., and Levine, R. P. (1965). Cytochrome f and plastocyanin: their sequence in the photosynthetic electron transport chain of *Chlamydomonas reinhardi*. *Proc. Natl. Acad. Sci. U.S.A.* 54, 1665–1669. doi:10.1073/pnas.54.6.1665

Greetham, D. (2014). Phenotype microarray technology and its application in industrial biotechnology. *Biotechnol. Lett.* 36, 1153–1160. doi:10.1007/s10529-014-1481-x

Harris, E. H. (2009). *The Chlamydomonas Sourcebook: Introduction to Chlamydomonas and Its Laboratory Use.* San Diego, CA: Academic Press.

Helliwell, K. E., Wheeler, G. L., Leptos, K. C., Goldstein, R. E., and Smith, A. G. (2011). Insights into the evolution of vitamin B12 auxotrophy from sequenced algal genomes. *Mol. Biol. Evol.* 28, 2921–2933. doi:10.1093/molbev/msr124

Holmes, D. S., and Bonnefoy, V. (2007). "Genetic and bioinformatic insights into iron and sulfur oxidation mechanisms of bioleaching organisms," in *Biomining*, eds D. E. Rawlings and D. B. Johnson (Berlin: Springer), 281–307.

Horton, P., Park, K.-J., Obayashi, T., Fujita, N., Harada, H., Adams-Collier, C. J., et al. (2007). WoLF PSORT: protein localization predictor. *Nucleic Acids Res.* 35, W585–W587. doi:10.1093/nar/gkm259

Kanehisa, M., Goto, S., Sato, Y., Kawashima, M., Furumichi, M., and Tanabe, M. (2014). Data, information, knowledge and principle: back to metabolism in KEGG. *Nucleic Acids Res.* 42, D199–D205. doi:10.1093/nar/gkt1076

Kirk, D. L., and Kirk, M. M. (1978). Carrier-mediated uptake of arginine and urea by *Chlamydomonas reinhardtii. Plant Physiol.* 61, 556–560. doi:10.1104/pp.61.4.549

Koskimaki, J. E., Blazier, A. S., Clarens, A. F., and Papin, J. A. (2013). Computational models of algae metabolism for industrial applications. *Ind. Biotechnol.* 9, 185–195. doi:10.1089/ind.2013.0012

Koussa, J., Chaiboonchoe, A., and Salehi-Ashtiani, K. (2014). Computational approaches for microalgal biofuel optimization: a review. *Biomed Res. Int.* 2014, 649453. doi:10.1155/2014/649453

Krupovic, M., Peixeiro, N., Bettstetter, M., Rachel, R., and Prangishvili, D. (2012). Archaeal tetrathionate hydrolase goes viral: secretion of a sulfur metabolism enzyme in the form of virus-like particles. *Appl. Environ. Microbiol.* 78, 5463–5465. doi:10.1128/AEM.01186-12

Manichaikul, A., Ghamsari, L., Hom, E. F. Y., Lin, C., Murray, R. R., Chang, R. L., et al. (2009). Metabolic network analysis integrated with transcript verification for sequenced genomes. *Nat. Methods* 6, 589–592. doi:10.1038/nmeth.1348

Matthew, T., Zhou, W., Rupprecht, J., Lim, L., Thomas-Hall, S. R., Doebbe, A., et al. (2009). The metabolome of *Chlamydomonas reinhardtii* following induction of anaerobic H2 production by sulfur depletion. *J. Biol. Chem.* 284, 23415–23425. doi:10.1074/jbc.M109.003541

May, P., Christian, J.-O., Kempa, S., and Walther, D. (2009). ChlamyCyc: an integrative systems biology database and web-portal for *Chlamydomonas reinhardtii. BMC Genomics* 10:209. doi:10.1186/1471-2164-10-209

Merchant, S. S., Prochnik, S. E., Vallon, O., Harris, E. H., Karpowicz, S. J., Witman, G. B., et al. (2007). The *Chlamydomonas* genome reveals the evolution of key animal and plant functions. *Science* 318, 245–250. doi:10.1126/science.1143609

Mettler, T., Mühlhaus, T., Hemme, D., Schöttler, M.-A., Rupprecht, J., Idoine, A., et al. (2014). Systems analysis of the response of photosynthesis, metabolism, and growth to an increase in irradiance in the photosynthetic model organism *Chlamydomonas reinhardtii. Plant Cell* 26, 2310–2350. doi:10.1105/tpc.114.124537

Monselise, E. I., Levkovitz, A., and Kost, D. (2014). Ultraviolet radiation induces stress in etiolated *Landoltia punctata*, as evidenced by the presence of alanine, a universal stress signal: a 15N NMR study. *Plant Biol.* doi:10.1111/plb.12198

Morgan, M. C., Boyette, M., Goforth, C., Sperry, K. V., and Greene, S. R. (2009). Comparison of the Biolog OmniLog Identification System and 16S ribosomal RNA gene sequencing for accuracy in identification of atypical bacteria of clinical origin. *J. Microbiol. Methods* 79, 336–343. doi:10.1016/j.mimet.2009.10.005

Munoz-Blanco, J., Hidalgo-Martinez, J., and Cardenas, J. (1990). Extracellular deamination of L-amino acids by *Chlamydomonas reinhardtii* cells. *Planta* 182, 194–198. doi:10.1007/BF00197110

Oberhardt, M. A., Palsson, B. Ø., and Papin, J. A. (2009). Applications of genome-scale metabolic reconstructions. *Mol. Syst. Biol.* 5, 320. doi:10.1038/msb.2009.77

Orth, J. D., Thiele, I., and Palsson, B. O. (2010). What is flux balance analysis? *Nat. Biotechnol.* 28, 245–248. doi:10.1038/nbt.1614

Príncipe, A., Jofré, E., Alvarez, F., and Mori, G. (2009). Role of a serine-type D-alanyl-D-alanine carboxypeptidase on the survival of *Ochrobactrum* sp. 11a under ionic and hyperosmotic stress. *FEMS Microbiol. Lett.* 295, 261–273. doi:10.1111/j.1574-6968.2009.01604.x

Pröschold, T., and Leliaert, F. (2007). "Systematics of the green algae: conflict of classic and modern approaches," in *Unravelling the Algae: The Past, Present, and Future of Algal Systematics*, eds J. Brodie and J. M. Lewis (Boca Raton, FL: CRC Press), 123–153.

Rohwerder, T., Gehrke, T., Kinzler, K., and Sand, W. (2003). Bioleaching review part A. *Appl. Microbiol. Biotechnol.* 63, 239–248. doi:10.1007/s00253-003-1448-7

Roth, J., Lawrence, J., and Bobik, T. (1996). Cobalamin (coenzyme B12): synthesis and biological significance. *Annu. Rev. Microbiol.* 50, 137–181. doi:10.1146/annurev.micro.50.1.137

Schellenberger, J., Que, R., Fleming, R. M., Thiele, I., Orth, J. D., Feist, A. M., et al. (2011). Quantitative prediction of cellular metabolism with constraint-based models: the COBRA Toolbox v2. 0. *Nat. Protoc.* 6, 1290–1307. doi:10.1038/nprot.2011.308

Schmidt, B. J., Lin-Schmidt, X., Chamberlin, A., Salehi-Ashtiani, K., and Papin, J. A. (2010). Metabolic systems analysis to advance algal biotechnology. *Biotechnol. J.* 5, 660–670. doi:10.1002/biot.201000129

Shu, L., and Hu, Z. (2012). Characterization and differential expression of microRNAs elicited by sulfur deprivation in *Chlamydomonas reinhardtii. BMC Genomics* 13:108. doi:10.1186/1471-2164-13-108

Takahashi, S., Okada, H., Abe, K., and Kera, Y. (2012). D-amino acid-induced expression of D-amino acid oxidase in the yeast *Schizosaccharomyces pombe. Curr. Microbiol.* 65, 764–769. doi:10.1007/s00284-012-0227-z

Tshikhudo, P., Nnzeru, R., and Mudau, F. (2013). Bacterial species identification getting easier. *Afr. J. Biotechnol.* 12, 5975–5982. doi:10.5897/AJB2013.12057

Vaas, L. A. I., Sikorski, J., Hofner, B., Fiebig, A., Buddruhs, N., Klenk, H.-P., et al. (2013). opm: an R package for analysing OmniLog® phenotype microarray data. *Bioinformatics* 29, 1823–1824. doi:10.1093/bioinformatics/btt291

Vaas, L. A. I., Sikorski, J., Michael, V., Göker, M., and Klenk, H.-P. (2012). Visualization and curve-parameter estimation strategies for efficient exploration of phenotype microarray kinetics. *PLoS ONE* 7:e34846. doi:10.1371/journal.pone.0034846

Varma, A., Boesch, B. W., and Palsson, B. O. (1993). Stoichiometric interpretation of Escherichia coli glucose catabolism under various oxygenation rates. *Appl. Environ. Microbiol.* 59, 2465–2473.

Veyel, D., Erban, A., Fehrle, I., Kopka, J., and Schroda, M. (2014). Rationales and approaches for studying metabolism in eukaryotic microalgae. *Metabolites* 4, 184–217. doi:10.3390/metabo4020184

Wase, N., Black, P. N., Stanley, B. A., and Dirusso, C. C. (2014). Integrated quantitative analysis of nitrogen stress response in *Chlamydomonas reinhardtii* using metabolite and protein profiling. *J. Proteome Res.* 13, 1373–1396. doi:10.1021/pr400952z

Winter, S. E., Thiennimitr, P., Winter, M. G., Butler, B. P., Huseby, D. L., Crawford, R. W., et al. (2010). Gut inflammation provides a respiratory electron acceptor for *Salmonella. Nature* 467, 426–429. doi:10.1038/nature09415

Yokoyama, T., Kan-No, N., Ogata, T., Kotaki, Y., Sato, M., and Nagahisa, E. (2003). Presence of free D-amino acids in microalgae. *Biosci. Biotechnol. Biochem.* 67, 388–392. doi:10.1271/bbb.67.388

Zuo, Z., Rong, Q., Chen, K., Yang, L., Chen, Z., Peng, K., et al. (2012). Study of amino acids as nitrogen source in *Chlamydomonas reinhardtii. Phycological Res.* 60, 161–168. doi:10.1111/j.1440-1835.2012.00646.x

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.

Received: 22 October 2014; paper pending published: 15 November 2014; accepted: 24 November 2014; published online: 10 December 2014.

Citation: Chaiboonchoe A, Dohai BS, Cai H, Nelson DR, Jijakli K and Salehi-Ashtiani K (2014) Microalgal metabolic network model refinement through high-throughput functional metabolic profiling. Front. Bioeng. Biotechnol. ***2****:68. doi: 10.3389/fbioe.2014.00068*

This article was submitted to Systems Biology, a section of the journal Frontiers in Bioengineering and Biotechnology.

frontiers in
BIOENGINEERING AND BIOTECHNOLOGY

METHODS ARTICLE
published: 26 January 2015
doi: 10.3389/fbioe.2014.00091

Integrative analysis of metabolic models – from structure to dynamics

Anja Hartmann[1]* and Falk Schreiber[2,3]

[1] Leibniz Institute of Plant Genetics and Crop Plant Research (IPK), Gatersleben, Germany
[2] Monash University, Melbourne, VIC, Australia
[3] Martin-Luther-University Halle-Wittenberg, Halle, Germany

Edited by:
Daniel Machado, University of Minho, Portugal

Reviewed by:
Rafael Costa, Instituto de Engenharia de Sistemas e Computadores, Investigação e Desenvolvimento em Lisboa (INESC-ID/IST), Portugal
Ina Koch, Goethe University Frankfurt am Main, Germany

***Correspondence:**
Anja Hartmann, Leibniz Institute of Plant Genetics and Crop Plant Research (IPK), Corrensstr. 3, OT Gatersleben, Stadt Seeland 06466, Germany
e-mail: hartmann@ipk-gatersleben.de

The characterization of biological systems with respect to their behavior and functionality based on versatile biochemical interactions is a major challenge. To understand these complex mechanisms at systems level modeling approaches are investigated. Different modeling formalisms allow metabolic models to be analyzed depending on the question to be solved, the biochemical knowledge and the availability of experimental data. Here, we describe a method for an integrative analysis of the structure and dynamics represented by qualitative and quantitative metabolic models. Using various formalisms, the metabolic model is analyzed from different perspectives. Determined structural and dynamic properties are visualized in the context of the metabolic model. Interaction techniques allow the exploration and visual analysis thereby leading to a broader understanding of the behavior and functionality of the underlying biological system. The System Biology Metabolic Model Framework (*SBM*^{2} – Framework) implements the developed method and, as an example, is applied for the integrative analysis of the crop plant potato.

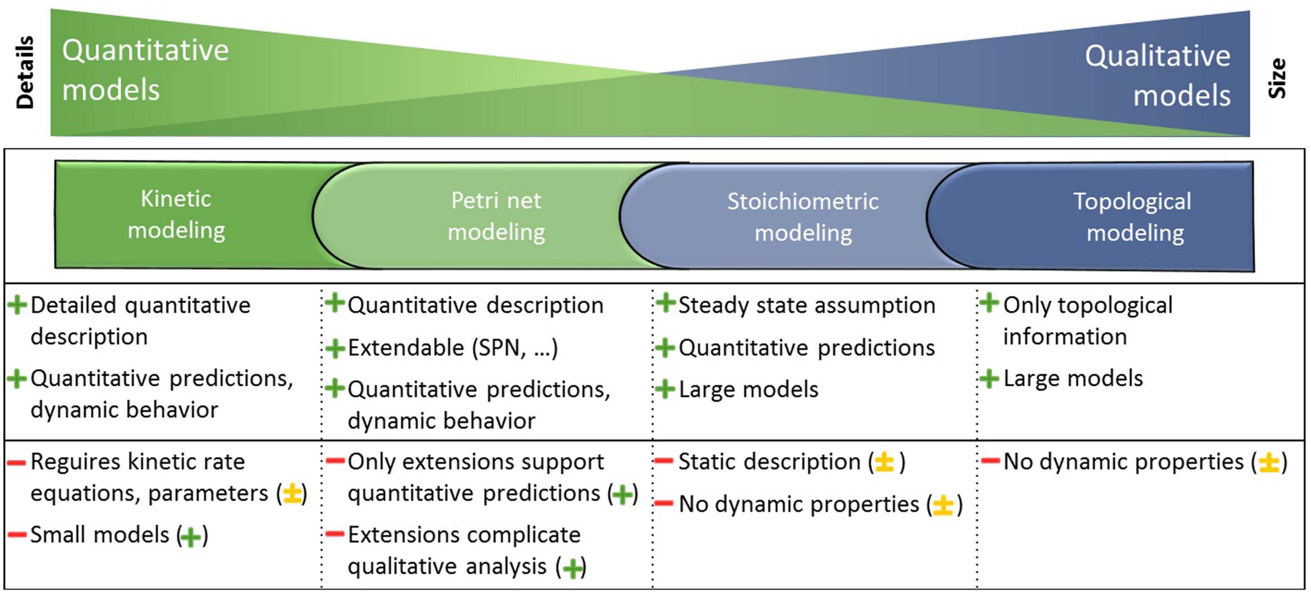
Keywords: metabolic modeling, integrative analysis, kinetic analysis, flux balance analysis, petri net analysis, topological analysis

1. INTRODUCTION

Metabolic models have been reconstructed for an increasing number of organisms to understand complex biochemical processes. At least 54 bacterial, 6 archaeal, and 16 eukaryotic reconstructions are available to-date while many others are under development (Xu et al., 2013). In addition, resources such as Path2Models (Büchel et al., 2013) provide draft models for a large number of organisms. Such metabolic models are composed of biochemical reactions and associated experimental parameters of the biological system under investigation. Different metabolic models can be reconstructed depending upon the completeness of knowledge about the detailed interaction mechanisms in a biological system. The metabolism is thereby roughly represented in large and mostly qualitative models and smaller, but more quantitative models (Steuer and Junker, 2008). Different model sizes and knowledge details allow the structural and dynamic properties to be analyzed using different modeling formalisms. For further details on modeling formalisms in Systems Biology the reader is referred to (Machado et al., 2011). Several modeling formalisms entail different analysis techniques facilitating the investigation of a metabolic model from different perspectives and thus, revealing complementary insights.

A couple of review papers evaluated modeling formalisms (Wiechert, 2002; Steuer and Junker, 2008; Hübner et al., 2011; Koch et al., 2011; Machado et al., 2011; Pfau et al., 2011; Dandekar et al., 2012) and revealed among others kinetic, Petri net, stoichiometric, and topological modeling methods as well-established. The strengths and weaknesses of each formalism are summarized in **Figure 1**.

Kinetic modeling using ordinary differential equations (ODEs) includes detailed quantitative descriptions on the biochemical processes and therefore requires often difficult to obtain kinetic rate equations and parameters. Due to this, kinetic modeling is generally limited to smaller models, but leads to quantitative predictions and reveals dynamic behavior of the underlying biological system (Resat et al., 2009). Petri net modeling is powerful due to several Petri net extensions for qualitative and quantitative analysis. The stochastic effects involved in quantitative predictions and system dynamics can be accounted for by using, for example, the stochastic Petri net (SPN) simulation. However, these extensions complicate the qualitative analysis (Baldan et al., 2010). Stoichiometric modeling using optimization-based analysis such as flux balance analysis (FBA) (Orth et al., 2010) allows for quantitative predictions due to the steady-state assumption. A static description of the biochemical processes is therefore sufficient when including stoichiometric, thermodynamic, and enzyme capacity constraints. Thus, stoichiometric modeling is applicable for large models, but is limited in revealing the dynamic behavior of the underlying biological system (Lewis et al., 2012). Topological modeling considers only the topological information of models (not limited in model size) and can identify structures and robustness against disturbances. Using, for example, centrality analysis (Koschützki and Schreiber, 2008) different importance concepts provide insights into key elements based on metabolite or reaction graphs (Steuer and Junker, 2008).



Kinetic modeling	Petri net modeling	Stoichiometric modeling	Topological modeling
+ Detailed quantitative description + Quantitative predictions, dynamic behavior	+ Quantitative description + Extendable (SPN, ...) + Quantitative predictions, dynamic behavior	+ Steady state assumption + Quantitative predictions + Large models	+ Only topological information + Large models
− Reguires kinetic rate equations, parameters (±) − Small models (+)	− Only extensions support quantitative predictions (+) − Extensions complicate qualitative analysis (+)	− Static description (±) − No dynamic properties (±)	− No dynamic properties (±)

FIGURE 1 | Metabolism is represented in large and mostly qualitative models and smaller, but more quantitative models. Different modeling formalisms that depend on the completeness of knowledge about the detailed interaction mechanisms are utilized to gain knowledge on the underlying biological system. Each modeling formalism is applied to different models as indicated by the corresponding colors and possesses strengths (+) and weaknesses (−). The integration of the independent modeling formalisms mitigates the weaknesses (−) and leads to the potential of the integrated analysis indicated in parentheses (major + or minor ± improvement, explanations are given in the Results and Discussion section). Adapted from Steuer and Junker (2008).

Some of the introduced metabolic modeling formalisms are already investigated in different approaches to analyze metabolic models at the system level and to overcome problems due to the lack of experimental data. Described methods either extent qualitative models with obtained analysis results to investigate a follow-up quantitative analysis, or models are reduced to assign less data for quantitative analysis. In most cases, such as Birch et al. (2014) and Chowdhury et al. (2014), the stoichiometric formalism FBA is used to obtain flux distributions, which are utilized to derive ODEs for kinetic analysis (Resat et al., 2009). Methods using the Petri net formalism for model reduction to integrate less data for kinetic analysis are described by Chen et al. (2011), Gilbert and Heiner (2006), and Koch and Heiner (2008). An advanced method is presented by Machado et al. (2010) whereby Petri net formalism is applied to integrate both of the aforementioned methods for model reduction and a follow-up kinetic analysis. Grafahrend-Belau et al. (2013) combined overview kinetic models (household models) with FBA toward a quasi-dynamic FBA. Heiner et al. (2012) and Nagasaki et al. (2010) propose a unifying Petri net framework comprised of a family of related Petri net types. In this approach qualitative, stochastic and continuous Petri net analyses are conducted by converting different Petri net types into each other.

Here, we introduce an integrated approach, which complements the presented approaches through a formalization leading to a standardized, transformable, and extensible abstraction of metabolism. This method allows the investigated metabolic models to be integrated, utilizing different well-established modeling formalisms and at the same time maintaining a standardized visualization. Moreover, the integration of analysis results with corresponding elements of the metabolic model leads to a combination of model structure and model dynamics. Several interaction techniques support the exploration and interpretation of the gained analysis results to provide a comprehensive understanding of the underlying biological system.

2. MATERIALS AND METHODS

In general, metabolic models are networks consisting of different elements such as metabolites and reactions with relations between these elements and additional attributes. Thus, a suitable data structure for metabolic models is a graph. Dependent upon the modeling formalism, graphs with different structure and attributes are able to represent kinetic, Petri net, stoichiometric, or topological models. Each of these graphs contains nodes (metabolites and/or reactions), which are related to each other through edges.

Following the concept of generalization, different *specific graphs* representing qualitative and quantitative metabolic models (**Figure 2C**) are generalized into a *unified graph* (**Figure 2A**). This concept allows a standard graphical representation to be maintained (**Figure 2B**) and additionally, to transform the *unified graph* into *specific graphs* to apply different modeling formalisms. Some formalisms utilize a reduced structure and attribute set of the *unified graph* to perform analyses (this will be described in detail in the Transformation Section). Using our method, the analysis results from different formalisms are visualized in the context of the metabolic model through data assignment functions (**Figure 2D**). Thus, the underlying

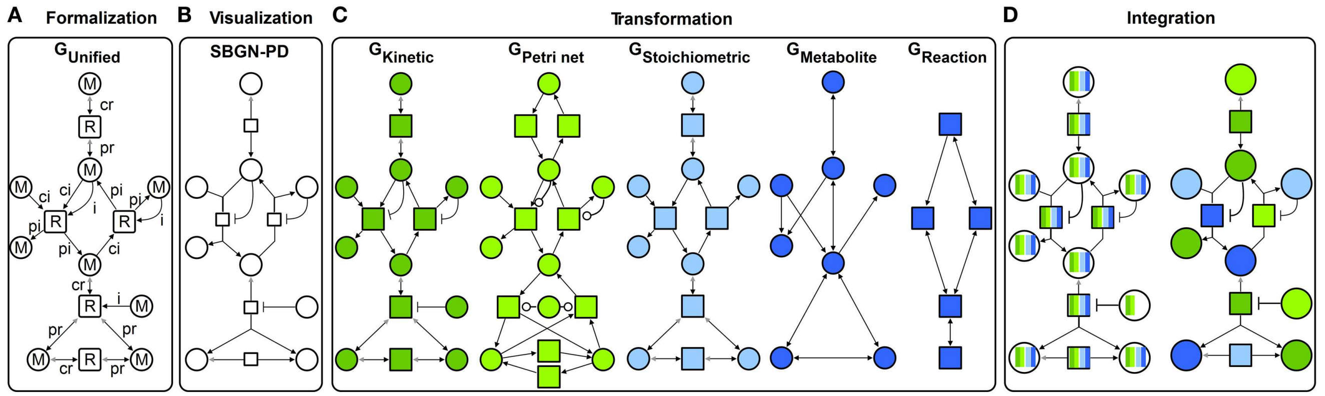


FIGURE 2 | Concept for an integrative analysis of metabolic models including: **(A)** formalization (G_{Unified} with *M* metabolite and *R* reaction nodes, different edge types: *ci* consumption irreversible, *cr* consumption reversible, *pi* production irreversible, *pr* production reversible, and *i* inhibition), **(B)** visualization in *SBGN-PD*, **(C)** transformation in different *specific graphs*: G_{Kinetic} dark green, $G_{\text{Petri net}}$ light green, $G_{\text{Stoichiometric}}$ light blue, and two topological graphs $G_{\text{Metabolite}}$ and G_{Reaction} dark blue, and **(D)** integration of different analysis results (colors represent results from different analysis performed using *specific graphs*).

biological system is characterized from different perspectives providing complementary insights. Using interaction techniques, the subsequent visual analysis is conducted. Furthermore, analysis results can be integrated in other formalisms to constrain this analysis and thereby make them either feasible or more precise.

The following sections introduce the concept depicted in **Figure 2** in detail.

2.1. FORMALIZATION

With the aim to formally represent qualitative and quantitative metabolic models a directed, attributed, bipartite graph (called the *unified graph*) is defined as follows.

Definition 2.1 (unified graph). The *unified graph* $G_{\text{Unified}} = (M, R, E, A)$ is a directed, attributed, bipartite graph consisting of two finite, non-empty sets M of metabolites and R of reactions, whereby both sets are disjoint $M \cap R = \varnothing$. Other finite sets are directed edges $E \subseteq (M \times R) \cup (R \times M)$ and attributes $A = \{type, stoichiometry, localization, label, concentration, capacity, rate, boundaries, objective\ function\}$, which are assigned to nodes and edges using the following functions:

- *type*: $E \rightarrow \{ci, pi, cr, pr, i\}$ is a function, which assigns a type to each edge (*ci* consumption irreversible, *pi* production irreversible, *cr* consumption reversible, *pr* production reversible, or *i* inhibition). A directed edge from a metabolite to a reaction is of type *ci*, *cr*, or *i* [i.e., $\forall e \in (M \times R)$: $type(e) = ci \vee type(e) = cr \vee type(e) = i$] and a directed edge from a reaction to a metabolite is of type *pi* or *pr* [i.e., $\forall e \in (R \times M)$: $type(e) = pi \vee type(e) = pr$]. To easily distinguish between reversible and irreversible edges, reversible edges are illustrated using a double-headed arrow, with the black arrow-head denoting the main direction from substrate (consumed metabolite) to product (produced metabolite) of a reaction.
- *stoichiometry*: $E' \rightarrow \mathbb{R}_{>0}$ is a function, which assigns a positive real number greater than 0 to each edge of type *ci*, *cr*, *pi*, or *pr* out of the set $E' = \{e \in E | \neg (type(e) = i)\}$.
- *label*: $M \cup R \rightarrow \Sigma^*$ is a function, which assigns a word over the alphabet to each metabolite and each reaction.
- *localization*: $M \rightarrow \Sigma^*$ is a function, which assigns a word over the alphabet to each metabolite.
- *capacity*: $M \rightarrow \mathbb{R}_{\geq 0} \cup \{\infty\}$ is a function, which assigns a positive real number or infinity $\{\infty\}$ to each metabolite.
- *concentration*: $M \rightarrow \mathbb{R}_{\geq 0}$ is a function, which assigns a positive real number to each metabolite. Additionally, the concentration of a metabolite has to be less than or equal to the capacity of the metabolite, $\forall m \in M$: $concentration(m) \leq capacity(m)$.
- *rate*: $R \rightarrow \{\{h, j\}, h, j, \{\}\}$ is a function, which assigns a kinetic rate equation $j \in J$, whereby J is a set of all kinetic rate equations or a positive real number (stochastic rate) $h \in \mathbb{R}_{\geq 0}$ or the empty set to each reaction.
- *boundaries*: $R \rightarrow (lower, upper)$, with $lower, upper \in \mathbb{R}_{\geq 0}$, and $lower \leq upper$ is a function, which assigns an ordered pair of positive real numbers to each reaction, whereby the lower bound has to be smaller than or equal to the upper bound.
- *objective function*: $R \rightarrow \{0, 1\}$, with $\forall r, r' \in R$: $objective\ function\ (r) = 1 \wedge objective\ function\ (r') = 1 \Rightarrow r = r'$, is a function, which assigns 0 or 1 to each reaction, whereby only one reaction receives the value 1 (for optimization).

Furthermore, the following requirements must be fulfilled:

For all reactions $r \in R$ applies: (1) there exists at least one incoming and one outgoing edge (whereby the incoming edge is not of type *i*) and (2) if one incoming or outgoing edge is reversible (irreversible) than all incoming and outgoing edges are reversible (irreversible). With this rule a reaction is either connected to reversible edges or irreversible edges but not a combination of them.

Between a metabolite $m \in M$ and a reaction $r \in R$ there are at most two edges $e, e' \in E$ of different types. If two edges e and e' connect m with r the type of e is *ci* and the type of e' is *i*. This case describes a substrate inhibition at high substrate concentrations, whereby a metabolite is substrate and inhibitor at the same time.

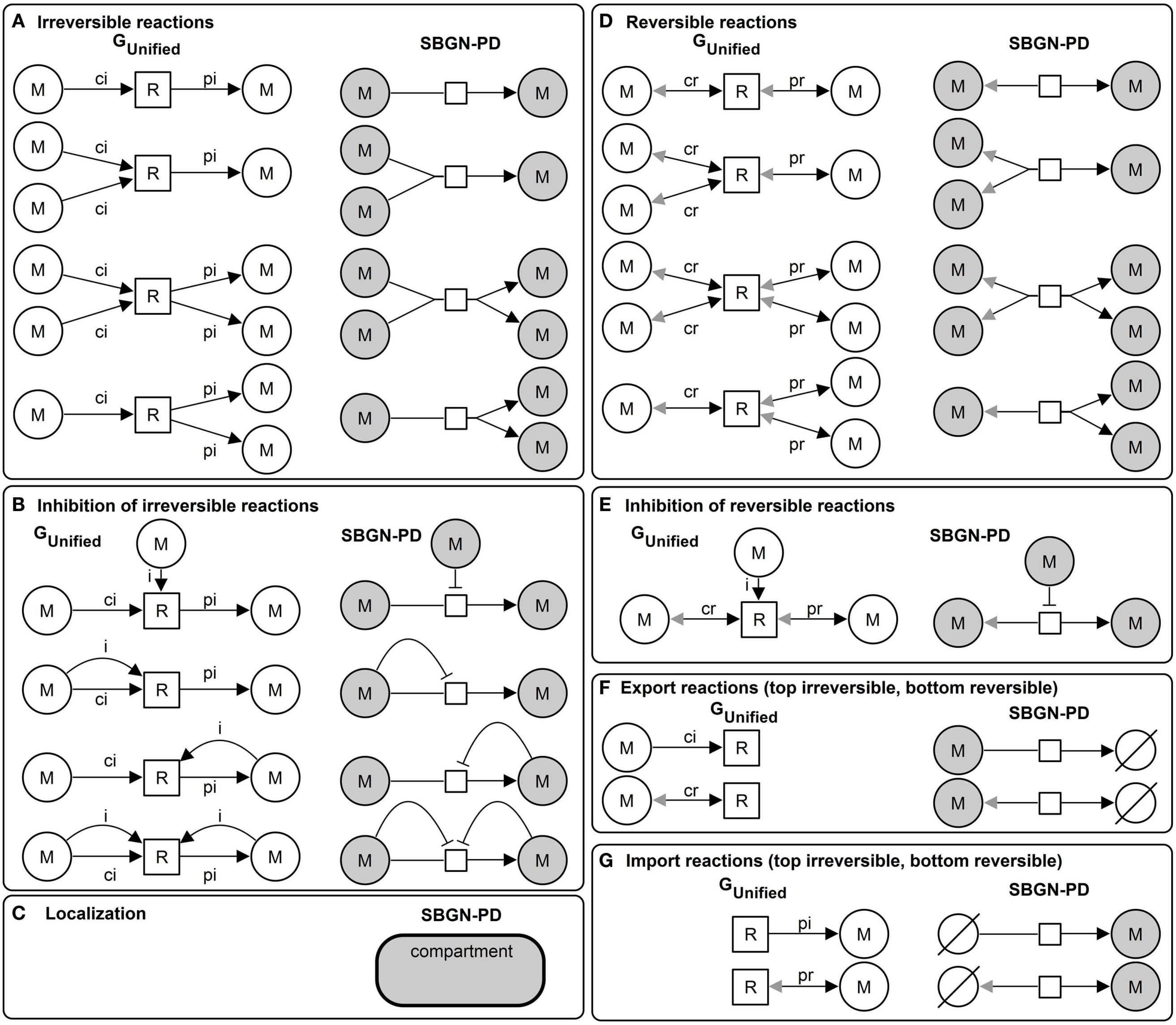

A Irreversible reactions
G_Unified
SBGN-PD
M
R
ci
pi
B Inhibition of irreversible reactions
i
C Localization
compartment
D Reversible reactions
cr
pr
E Inhibition of reversible reactions
F Export reactions (top irreversible, bottom reversible)
G Import reactions (top irreversible, bottom reversible)

FIGURE 3 | Basic elements of the *unified graph* (left) and the corresponding *SBGN-PD* visualization (right): **(A)** irreversible reactions, **(B)** inhibition of irreversible reactions, **(C)** localization (compartment) of metabolites (samecolor), **(D)** reversible reactions, **(E)** inhibition of reversible reactions, **(F)** export reactions (top irreversible and bottom reversible), and **(G)** import reactions (top irreversible, bottom reversible).

If one edge e connects r with m and another edge e' connects m with r the type of e is pi and the type of e' is i. In this case, a product inhibition is modeled with a metabolite as product and at the same time inhibitor of a reaction.

An explicit formulation of both cases for reversible reactions is not needed because the reaction mechanisms already provide implicit substrate- and product inhibition.

Moreover, the following sets are defined to simplify the transformation of the *unified graph* into *specific graphs* for analysis. The edge set E is composed of three subsets, $E = E_i \cup E_{ir} \cup E_r$. The subset of inhibitory edges is $E_i = \{e \in E | type(e) = i\}$, the subset of irreversible edges is $E_{ir} = \{e \in E | type(e) = ci \vee type(e) = pi\}$ and the subset of reversible edges is $E_r = \{e \in E | type(e) = cr \vee type(e) = pr\}$. The set of metabolites M consists of a subset of metabolites M_{cp}, which are either consumed or produced in reactions $M_{cp} = \{m \in M | \exists r \in R\colon (m, r) \in E_r \vee (m, r) \in E_{ir}\} \cup \{m' \in M | \exists r \in R\colon (r, m') \in E_r \vee (r, m') \in E_{ir}\}$.

To assign analysis results to nodes and edges of the *unified graph*, data assignment functions that integrate calculated structural and dynamic data are used (this will be described in detail in the Transformation section).

Due to the definition of the *unified graph* with a rich attribute set qualitative and quantitative metabolic models can be represented and additionally visualized using standards. **Figure 3** illustrates the basic elements of the *unified graph* and the corresponding visualization in *SBGN-PD*.

2.2. VISUALIZATION

In order to derive a standardized graphical representation of the *unified graph* the Systems Biology Graphical Notation (Le Novère et al., 2009) (*SBGN*) is utilized. *SBGN* has been developed to interpret biological models easily without the need for extensive descriptions using three sub-languages. *SBGN-PD* (Moodie et al., 2011) is the *Process Description* sub-language visualizing the temporal dependencies of biological interactions in detail and is thus suited for the metabolic models encoded in the *unified graph*.

The translation of the *unified graph* in a *SBGN-PD* visualization is based on the following schema. All elements of the metabolite set $m \in M$ (reaction set $r \in R$) are visualized using *simple chemicals* ∈ *entity pool nodes* (*process* ∈ *process nodes*). All elements of the edge set $e \in E$ are visualized using arcs of the set *connecting arcs* based on the assigned type. Edges of type *ci* are visualized using *consumption arc*, *pi* using *production arc*, *cr* using *production arc* in the opposite direction, *pr* using *production arc* and *i* using *inhibition arc*, respectively.

The edge attribute *stoichiometry* is visualized using *cardinality* and the metabolite attribute *localization* is visualized using *compartment*, which is a container for metabolites defined for this location. The localization of reactions is independent of a *compartment*, hence, a reaction could be located within, outside or on top of the border of a *compartment*. Import or export reactions in *SBGN-PD* are defined using the additional symbol *source and sink* ∈ *entity pool nodes*, see **Figure 3**.

Furthermore, interaction techniques allow the exploration and subsequent visual analysis leading to a broader understanding of the behavior and functionality of the underlying biological system (which will be described in detail in the Results and Discussion section).

2.3. TRANSFORMATION

Overall, five transformations from the *unified graph* (G_{Unified}) into the *specific graphs* (G_{Kinetic}, $G_{\text{Petri net}}$, $G_{\text{Stoichiometric}}$, $G_{\text{Metabolite}}$, G_{Reaction}) have to be performed as a prerequisite to analyze a metabolic model using different modeling formalisms. The different models, modeling formalisms and the transformation from G_{Unified} into $G_{\text{Stoichiometric}}$ are described in the following. The transformations from G_{Unified} into G_{Kinetic}, $G_{\text{Petri net}}$, and into both of the topological graphs $G_{\text{Metabolite}}$, G_{Reaction} are defined in the Supplementary Material.

2.3.1. Kinetic model

A kinetic metabolic model (ODE model) consists of a structural description of relations between metabolites and reactions and is extended with detailed kinetic data including rate equations, metabolite concentrations, and additional kinetic parameters. The kinetic model is represented by the *kinetic graph* (G_{Kinetic}), which is transformed from the *unified graph* (G_{Unified}), see **Figure 2C** and for details Definition 1.1 in Supplementary Material. This transformation results in no structural differences, but in a reduced attribute set.

To analyze the kinetic metabolic model its *kinetic graph* is converted in ODEs, which are numerically solved (Resat et al., 2009). Changes in metabolite concentrations and reaction rates over a period of time are obtained as the results of the analysis.

2.3.2. Petri net model

A Petri net metabolic model can be defined using different Petri net types. Here, we refer to extended qualitative place/transition Petri nets (*eP/T nets*) and extended quantitative stochastic Petri nets (*eSPNs*). The extension includes continuous tokens (to model metabolite concentrations), continuous arc weights (to model non-integer stoichiometry), continuous place capacities (to model limited resources), and inhibitor arcs (to model inhibition). An inhibition is modeled using an inhibitor arc from a place to a transition meaning that the transition can only fire if no token is on that place. The transition may only fire when the place is empty.

Both Petri net types share the same structure, but *eSPNs* are specialized by weights for the exponentially distributed random variable (firing time) assigned to transitions. For further details on Petri nets for modeling metabolic models the reader is referred to Baldan et al. (2010). The Petri net model is represented by the *Petri net graph* ($G_{\text{Petri net}}$), which is transformed from the *unified graph* (G_{Unified}), see **Figure 2C** and for details Definition 1.2 and Figure S1 in Supplementary Material. This transformation results in structural differences (reversible reactions are represented using a pair of irreversible reactions for both directions) and a reduced attribute set.

A Petri net metabolic model can be analyzed qualitatively or quantitatively. For the qualitative analysis, the *Petri net graph* is converted into a linear equation system, which can be solved to derive invariants describing main pathways (T-invariants) or metabolite conservation (P-invariants) of a metabolic model [more details in Murata (1989), Baldan et al. (2010), and Reisig (2013)]. Furthermore, all possible states are calculated using the reachability analysis and if the reachability graph cannot be constructed then the coverability graph is calculated instead (infinite state-space). The main purpose of the quantitative analysis (simulation) of a Petri net metabolic model is to include stochastic effects. The reactions can additionally be weighted with reaction rates to conduct a more constraint stochastic simulation revealing changes in metabolite concentrations over a number of simulation steps.

2.3.3. Stoichiometric model

Compared to both of the aforementioned models a stoichiometric model consists of stoichiometric reactions without quantities, such as metabolite concentrations, or reaction rates. Due to the steady-state assumption, the regulatory effects resulting from enzymes or inhibitors are neglected; see Orth et al. (2010) for more details.

Definition 2.2 (stoichiometric graph). The *unified graph* G_{Unified} is transformed in a directed, attributed, bipartite *stoichiometric graph* $G_{\text{Stoichiometric}} = (M_S, R_S, E_S, A_S)$ with a metabolite set $M_S = M_{cp}$, which is a subset of the set M in G_{Unified}. Metabolites with only inhibitory interactions to reactions are not considered. The reaction set in $G_{\text{Stoichiometric}}$ $R_S = R$ equals the reaction set R set in $G_{Unified}$ and the edge set in $G_{\text{Stoichiometric}}$ $E_S = E_{ir} \cup E_r$ is a subset of the set E in $G_{Unified}$. Edges of type i are excluded. The attribute set in $G_{\text{Stoichiometric}}$ $A_S \subseteq A$ is a subset of the set A in G_{Unified} with $A_S = \{type, stoichiometry, localization, label, boundaries, objective\ function\}$.

Figure 2C and for details Figure S4 in Supplementary Material depict the transformation of inhibited reactions from G_{Unified} into $G_{\text{Stoichiometric}}$ and thereby detailing the difference between both graphs. This transformation results in structural differences (no inhibitions) and a reduced attribute set. Thereby, all regulatory information and quantitative data are lost.

Using the *stoichiometric graph*, a metabolic model can be validated utilizing the *Dead-End* analysis or *Gap-Finding* analysis revealing blocked reactions or dead-end metabolites. To examine the flow of metabolites through a metabolic model the *stoichiometric graph* is converted into a system of mass balance equations at steady-state, which are solved by minimizing or maximizing an objective function. This optimization can be conducted using a linear optimization instead of a non-linear optimization to handle the problem of alternate optimal solutions. Applicable optimization-based methods are *FBA*, flux variability analysis (*FVA*), robustness analysis (*RA*), and knockout-analyses (*KA*) resulting in a flux distribution, minimal and maximal fluxes, sensitivity curves, and sensitivity values, respectively. For a detailed description of optimization-based methods the reader is referred to (Lewis et al., 2012).

2.3.4. *Topological models*

Metabolic models are analyzed according to topological properties in order to understand the importance of key elements, structure, and robustness against disturbances. Since the *metabolite graph* (nodes represent metabolites, edges reactions) and *reaction graph* (nodes represent reactions, edges metabolites) are predominantly used for topological analysis (Steuer and Junker, 2008) the *unified graph* G_{Unified} is transformed into both, see **Figure 2C** (For details see Definition 1.3 and Figure S2 in Supplementary Material for *metabolite graph* and Definition 1.4 and Figure S3 in Supplementary Material for *reaction graph*). This transformation results in structural differences (unipartite graphs) and a reduced attribute set. Thereby, all regulatory information and quantitative data are lost.

Topological analysis of the metabolic model based on its *metabolite graph* or *reaction graph* is conducted using the corresponding adjacency matrix. A shortest path analysis results in paths (subgraphs which could be the graph itself). Furthermore, centrality analysis with different centrality measures leads to a ranking of graph elements according to different importance concepts. For further details on different centrality measures the reader is referred to Koschützki and Schreiber (2008).

2.4. INTEGRATION

To integrate structural and dynamic analysis results in the *unified graph*, which have been computed using *specific graphs*, data assignment functions are applied. To focus on several analysis methods, we chose typical examples from a number of analysis methods comprised in the different modeling formalisms. Using these analysis methods, two sets of data types are generated: vectors of numeric values and graph elements, which are assigned to different graph elements of the *unified graph*, see **Table 1**.

Numeric values of the vector ($nv \in NV$) are assigned to elements of the *unified graph* (M metabolite, R reaction, and E edge) using the assignment function zn: $M, R, E \rightarrow NV$, whereby the vector could comprise numeric values (e.g., sensitivity values),

Table 1 | Summary of typical examples of analysis methods and corresponding results produced with different modeling formalisms grouped in data types, which will be assigned to different graph elements [metabolite nodes (*M*), reaction nodes (*R*), and edges (*E*)] of the *unified graph*.

Modeling formalisms	Typical examples of analysis methods		Analysis results	Data types	G_{Unified}		
					M	***R***	***E***
Kinetic modeling	Kinetic analysis		Metabolite concentrations, reaction rates over time	Vector of time dependent numeric values	x	x	
Petri net modeling	Invariant analysis		P- and T-invariants	Vector of numeric values	x[a]	x[a]	
	Reachability analysis		Reachability graph/coverability graph	Graph	x[a]	x[a]	
	Stochastic analysis		Metabolite concentrations, reaction rates over steps	Vector of step dependent numeric values	x[a]	x[a]	
Stoichiometric modeling	Stoichiometric analysis		Dead-ends	Nodes	x		
	Optimization-based analysis	Gap-finding	Gaps	Nodes	x		
		FBA	Flux distribution	Vector of numeric values			x
		RA	Sensitivity curve	Vector of flux dependent numeric values		x	
		KA	Sensitivity value	Vector of numeric values		x	
		FVA	Min/max flux values of reactions	Vector of numeric value pairs		x	
Topological modeling	Centrality analysis		Centrality values	Vector of numeric values	x	x	
	Shortest path		Shortest path	Graph	x[b]	x[b]	x[b]

[a] *Analysis results from forward and backward reactions of the Petri net are integrated into the corresponding reversible reactions in the unified graph.*

[b] *Analysis results from edges of the metabolite graph or reaction graph correspond to several edges and nodes in the unified graph.*

pairs of numeric values (e.g., min and max fluxes), and a set of time, step, and flux value dependent numeric values (e.g., metabolite concentrations over time, steps and sensitivity curves, respectively).

Another type of analysis results data are the elements of graphs, which are assigned to the *unified graph* using the assignment function *zg*: M, R, $E \rightarrow M_x$, R_x, E_x, whereby x can be replaced with P Petri net, S stoichiometric, K kinetic, M metabolite, or R reaction to define the *specific graphs*. As an example, *Gap-Finding* analysis results in a set of metabolites of the *stoichiometric graph*, which must be assigned to metabolites in the *unified graph* using *zg*: $M \rightarrow M_S$.

These assignment functions provide the basis for the visualization of the analysis results in the context of the metabolic model. Furthermore, interaction techniques such as *brushing* & *linking* and *animation* support the exploration, for example, of different Petri net invariants in the context of the metabolic model [for more details concerning interaction techniques see Von Landesberger et al. (2011)]. An integrated visualization by means of an application using the developed method is represented in the Results and Discussion section.

3. RESULTS AND DISCUSSION

In conclusion, the developed method allows previously separated well-established modeling formalisms to be combined into one application using one workflow, supported by interaction techniques and integrated visualizations in the context of the metabolic model. The method mitigates the weaknesses (−) of independent modeling formalisms as explained in the Introduction section and leads to major (+) or minor (±) improvements of an integrated analysis as already depicted in **Figure 1**.

In detail, using the integrated approach it is not required to define detailed kinetics to derive quantitative predictions and reveal dynamic behavior of the underlying biological system. Instead, using some parameters the Petri net simulation or stoichiometric modeling method FBA could be performed to approximate kinetic simulations. Thus, larger models are applicable in the integrated approach leading to analysis results, which could be again integrated to analyze the model further. Additionally, qualitative analysis can be conducted for extended Petri nets using another integrated formalism such as Dead-End analysis or centrality analysis. Quantitative predictions can be revealed for a qualitative model with a static description using stoichiometric analysis.

Hence, different modeling formalisms complement each other even through, overlaps between the introduced metabolic modeling formalisms exist. For example, the stoichiometric matrix used in the stoichiometric modeling formalism to derive mass balance equations corresponds to the incidence matrix of the Petri net formalism used to derive an equation system solved for, e.g., invariant analysis. In the case of structural analysis, both the stoichiometric and the Petri net formalism could be utilized to reveal, for example, *Dead-End* metabolites. Additionally, Petri net T-invariants correspond to flux modes, which could be directly calculated using the stoichiometric analysis method elementary flux modes (not presented here).

The described method is implemented as an Add-on for the *VANTED* system (Rohn et al., 2012), called the System Biology Metabolic Model Framework (*SBM*^{2} – Framework). It utilizes and extends *VANTED*s functionality for the interpretation of experimental data and for analyzing metabolic models with different modeling formalisms.

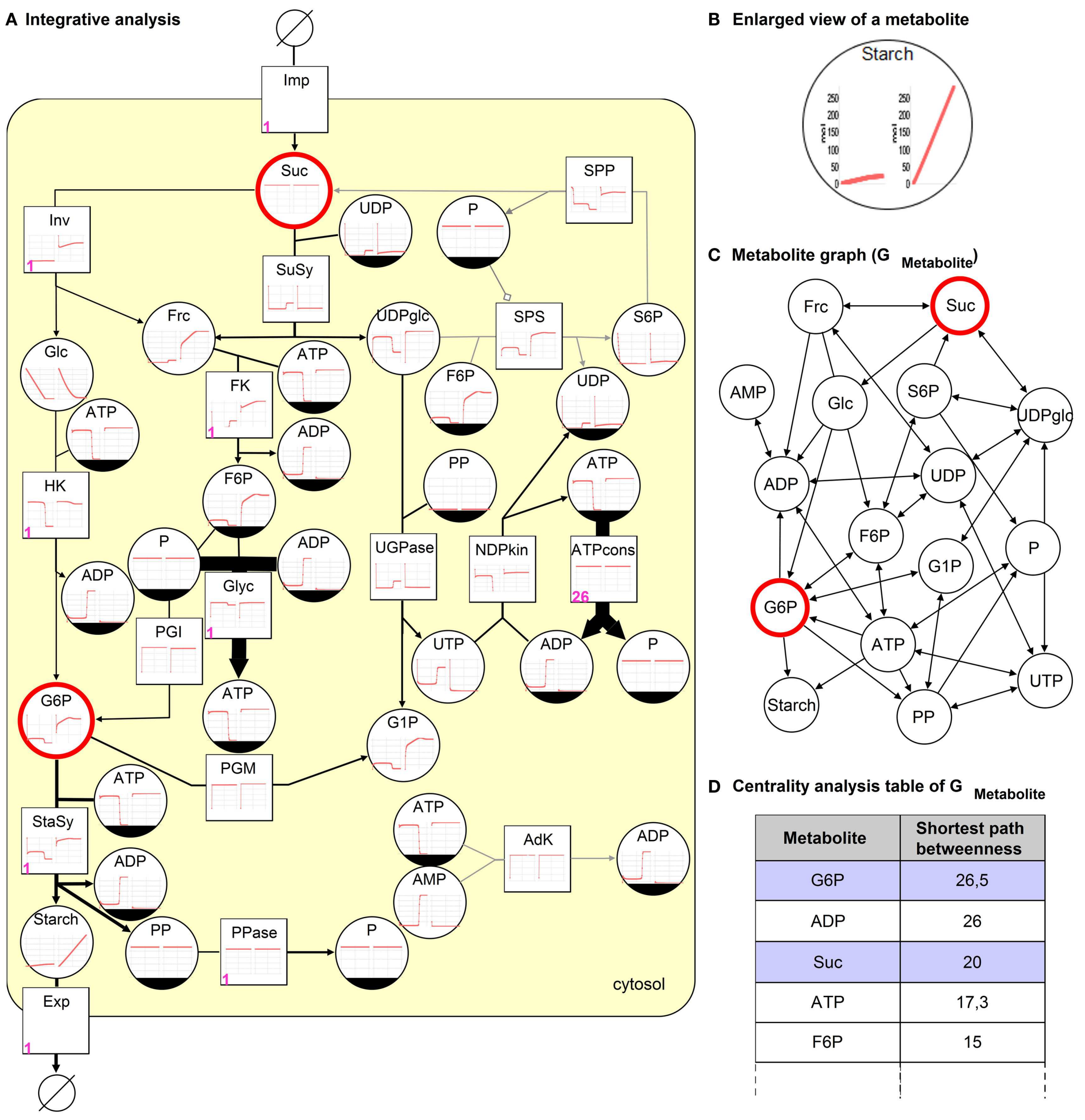
In order to characterize the metabolic functionality and behavior of the crop plant potato (*Solanum tuberosum*) an integrative analysis is performed using the described method. Due to its main component, starch in the potato tuber, potato is of great importance as food and in industry, for example, for the production of fuel. Therefore, a major aim of plant breeding is to improve the distribution of biomass within the plant in favor of harvestable plant parts. Based on the homogeneous tissue of the potato tuber the main flux of metabolites is from sucrose to starch (Geigenberger et al., 2004). The investigation of sucrose degradation can be conducted. Almost all genes of this pathway are already known and thus provide the basis for the reconstruction of a metabolic model of the potato tuber.

Using a kinetic model representing the sucrose breakdown in the developing potato tuber (Junker, 2004) the integrative analysis is performed and analysis results are shown in **Figure 4A**. The model comprises of 15 reactions and 17 metabolites located in the cytosol. Sucrose (*Suc*) is converted into hexose phosphates (e.g., glucose-6 phosphate, *G6P*) utilized in glycolysis (*Glyc*) and as precursors for starch synthase (*StaSy*). The pathways *Glyc*, starch biosynthesis, and energy consumption (ATP_{cons}) are modeled as summarized reactions. This is a necessary simplification to avoid unknown transport processes into additional compartments. To describe the environment the model is extended through sucrose import (*Imp*) and starch export reactions (*Exp*).

The kinetic analysis results in time-course diagrams converging toward a steady-state producing starch, which can be increased by an overexpression of the enzyme invertase (*Inv*) as described in Junker (2004). The consequence of the overexpression can be compared and visually analyzed to investigate both situations side by side in the model, see **Figures 4A,B**.

To perform a stochastic simulation the steady-state reaction rates generated by the kinetic analysis are used to weight the reactions of the eSPN. The stochastic simulation results in increasing and decreasing metabolite concentrations, which oscillate with different amplitudes (data not shown). The results indicate the production of starch and the utilization of reactions with different probabilities.

Additionally, the invariant analysis reveals beside 3 P-invariants (reflecting substance conservation) 19 T-invariants, which can be grouped in trivial and non-trivial T-invariants. Each of the seven trivial T-invariants corresponds to a reversible reaction. The non-trivial T-invariants can be differentiated in a group of nine representing the cleavage of sucrose by invertase and another group of three where the sucrose is cleaved by sucrose synthase. These T-invariants reflect the main processes that are pathways taking place in the metabolic model in reality (Koch et al., 2005). One of the T-invariants is illustrated in **Figure 4A** by adding numbers (firing counter) to the corresponding reactions. Sucrose is initially cleaved by invertase, leading to the production of hexose phosphates, which are metabolized in *Glyc* and starch biosynthesis.



Metabolite	Shortest path betweenness
G6P	26,5
ADP	26
Suc	20
ATP	17,3
F6P	15

FIGURE 4 | Integrative analysis of the sucrose breakdown in the potato tuber. **(A)** The kinetic analysis results in time-course diagrams of metabolites and reactions (left wild type, right overexpression of *Inv*), **(B)** enlarged view of both diagrams for metabolite starch. Petri net invariant analysis results in T-invariants, one is represented using numbers (firing counter, left lower corner in pink) assigned to reactions. The steady-state flux distribution resulting from FBA optimized for maximization of starch biosynthesis is depicted as edge thickness (gray edge indicates 0 flux). **(C)** The topological analysis (shortest path betweenness centrality analysis) of the *metabolite graph* results in a **(D)** ranked table. Two metabolites are selected in the table (blue), which correspond to the highlighted (red) nodes in **(A,C)**.

The stoichiometric analysis (irrespective regulatory processes), using only three steady-state reaction rates ($Inv = 0.16\,\mu M/FW/s$, $SuSy = 4.89\,\mu M/FW/s$, $ATPcons = 100\,\mu M/FW/s$) to constrain the fluxes for these reactions, results in a flux distribution, which is comparable to the kinetic analysis results. In **Figure 4A**, the edge thickness corresponds to flux values. The flux through the starch biosynthesis reaction with $6.42\,\mu M/FW/s$ is equal to the one of the kinetic analysis. Additionally, the reaction *AdK* is not utilized as can be seen in results of the kinetic and Petri net analysis.

Using the *metabolite graph*, see **Figure 4C**, the structure of the potato model is investigated. To identify important metabolites that occur on the shortest paths between two nodes in a ranked way the *shortest path betweenness (SPB)* centrality analysis is conducted. As a result, the table in **Figure 4D** illustrates *Suc* and *G6P*, which are selected to be highlighted in **Figures 4A,C**. Both metabolites are very important in the model, indicating that without these metabolites the reactions of starch biosynthesis and *Glyc* could not be processed.

In summary, using the integrative analysis allows different modeling formalisms to be investigated in one workflow. An integrated and interactive visualization of the analysis results leads to an advantage over the use of each modeling formalism independently. This helps to compare analysis results from different formalisms within one metabolic model and allows for the investigation of analysis results from one formalism in another, as mentioned in the use case.

4. CONCLUSION

We described a method, which is able to bring together different metabolic modeling formalisms. The integration is realized by a *unified graph*, enabling graph transformations, and a visualization in a standardized and formalized way. The *unified graph* supports user interaction and thereby allows different analysis results to be explored in the context of the metabolic model. The application reveals structural and dynamic properties of the crop plant potato utilizing the integrative analysis. The method has been implemented as an extension of the *VANTED* system and could also be applied to other model types, but we have focused here on metabolic models as an application area.

Combining different modeling formalisms opens many possibilities for future research. Additional analysis algorithms can be added to study metabolic models in more detail. We plan to extend the method for different types of models such as gene regulatory models to investigate further cellular processes. This extension requires the adaptation of the *unified graph*, adding of appropriate modeling formalisms, and corresponding transformations. Furthermore, the visualization has to be adapted to represent different types of models in *SBGN* using, for example, the sub-language *SBGN-AF* for gene regulatory models.

AUTHOR CONTRIBUTIONS

Anja Hartmann developed the theoretical framework, the use case, and implemented the *SBM*^{2} – Framework software. Falk Schreiber supervised the project and gave conceptual advice. Both authors wrote the manuscript.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at http://www.frontiersin.org/Journal/10.3389/fbioe.2014.00091/abstract

REFERENCES

Baldan, P., Cocco, N., Marin, A., and Simeoni, M. (2010). Petri nets for modelling metabolic pathways: a survey. *Nat. Comput.* 9, 955–989. doi:10.1007/s11047-010-9180-6

Birch, E. W., Udell, M., and Covert, M. W. (2014). Incorporation of flexible objectives and time-linked simulation with flux balance analysis. *J. Theor. Biol.* 345, 12–21. doi:10.1016/j.jtbi.2013.12.009

Büchel, F., Rodriguez, N., Swainston, N., Wrzodek, C., Czauderna, T., Keller, R., 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

Chen, M., Hariharaputran, S., Hofestädt, R., Kormeier, B., and Spangardt, S. (2011). Petri net models for the semi-automatic construction of large scale biological networks. *Nat. Comput.* 10, 1077–1097. doi:10.1007/s11047-009-9151-y

Chowdhury, A., Zomorrodi, A. R., and Maranas, C. D. (2014). k-OptForce: integrating kinetics with flux balance analysis for strain design. *PLoS Comput. Biol.* 10:e1003487. doi:10.1371/journal.pcbi.1003487

Dandekar, T., Fieselmann, A., Majeed, S., and Ahmed, Z. (2012). Software applications toward quantitative metabolic flux analysis and modeling. *Brief. Bioinformatics* 15, 91–107. doi:10.1093/bib/bbs065

Geigenberger, P., Stitt, M., and Fernie, A. R. (2004). Metabolic control analysis and regulation of the conversion of sucrose to starch in growing potato tubers. *Plant Cell Environ.* 27, 655–673. doi:10.1111/j.1365-3040.2004.01183.x

Gilbert, D., and Heiner, M. (2006). "From Petri nets to differential equations – an integrative approach for biochemical network analysis," in *ICATPN, Volume 4024 of Lecture Notes in Computer Science*, eds S. Donatelli and P. S. Thiagarajan (Berlin: Springer), 181–200.

Grafahrend-Belau, E., Junker, A., Eschenröder, A., Müller, J., Schreiber, F., and Junker, B. H. (2013). Multiscale metabolic modeling: dynamic flux balance analysis on a whole-plant scale. *Plant Physiol.* 163, 637–647. doi:10.1104/pp.113.224006

Heiner, M., Herajy, M., Liu, F., Rohr, C., and Schwarick, M. (2012). "Snoopy – a unifying Petri net tool," in *Application and Theory of Petri Nets, Volume 7347 of Lecture Notes in Computer Science*, eds S. Haddad and L. Pomello (Berlin: Springer), 398–407.

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

Junker, B. H. (2004). *Sucrose Breakdown in the Potato Tuber.* Dissertation, Faculty of Science, Potsdam: University of Potsdam.

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. (2011). *Modeling in Systems Biology: The Petri net Approach.* New York, NY: Springer, 16.

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.

Le Novère, N., Hucka, M., Mi, H., Moodie, S., Schreiber, F., Sorokin, A., et al. (2009). The systems biology graphical notation. *Nat. Biotechnol.* 27, 735–741. doi:10.1038/nbt0909-864d

Lewis, N. E., Nagarajan, H., and Palsson, B. O. (2012). Constraining the metabolic genotype-phenotype relationship using a phylogeny of in silico methods. *Nat. Rev. Microbiol.* 77, 541–580. doi:10.1038/nrmicro2737

Machado, D., Costa, R., Rocha, M., Ferreira, E. C., Tidor, B., and Rocha, I. (2011). Modeling formalisms in systems biology. *AMB Express* 1, 45–58. doi:10.1186/2191-0855-1-45

Machado, D., Costa, R. S., Rocha, M., Rocha, I., Tidor, B., and Ferreira, E. C. (2010). "Model transformation of metabolic networks using a Petri net based framework," in *ACSD/Petri Nets Workshops, Volume 827 of CEUR Workshop Proceedings*, eds S. Donatelli, J. Kleijn, R. J. Machado, and J. M. Fernandes (Braga: CEUR-WS.org), 103–117.

Moodie, S., Novère, N. L., Demir, E., Mi, H., and Schreiber, F. (2011). Systems biology graphical notation: process description language level 1. *Nat. Proc.* doi:10.1038/npre.2011.3721.4

Murata, T. (1989). Petri nets: properties, analysis and applications. *Proc. IEEE* 10, 291–291.

Nagasaki, M., Saito, A., Jeong, E., Li, C., Kojima, K., Ikeda, E., et al. (2010). Cell illustrator 4.0: a computational platform for systems biology. *In silico Biol.* 10, 5–26.

Orth, J. D., Thiele, I., and Palsson, B. O. (2010). What is flux balance analysis? *Nat. Biotechnol.* 28, 245–248. doi:10.1038/nbt.1614

Pfau, T., Christian, N., and Ebenhöh, O. (2011). Systems approaches to modelling pathways and networks. *Brief. Funct. Genomics* 10, 266–279. doi:10.1093/bfgp/elr022

Reisig, W. (2013). *Understanding Petri Nets – Modeling Techniques, Analysis Methods, Case Studies*. Berlin: Springer.

Resat, H., Petzold, L., and Pettigrew, M. F. (2009). "Kinetic modeling of biological systems," in *Computational Systems Biology., Volume 541 of Methods in Molecular Biology*, eds R. Ireton, K. Montgomery, R. Bumgarner, R. Samudrala, and J. McDermott (New York: Humana Press), 311–335.

Rohn, H., Junker, A., Hartmann, A., Grafahrend-Belau, E., Treutler, H., Klapperstück, M., et al. (2012). Vanted v2: a framework for systems biology applications. *BMC Syst. Biol.* 6:139. doi:10.1186/1752-0509-6-139

Steuer, R., and Junker, B. H. (2008). "Computational models of metabolism: stability and regulation in metabolic networks," in *Advances in Chemical Physics*, Vol. 142, ed. S. A. Rice (Hoboken: John Wiley and Sons, Inc), 105–251. doi:10.1002/9780470475935.ch3

Von Landesberger, T., Kuijper, A., Schreck, T., Kohlhammer, J., van Wijk, J., Fekete, J.-D., et al. (2011). Visual analysis of large graphs: state-of-the-art and future research challenges. *Comput. Graph. Forum* 30, 1719–1749. doi:10.1111/j.1467-8659.2011.01898.x

Wiechert, W. (2002). Modeling and simulation: tools for metabolic engineering. *J. Biotechnol.* 94, 37–63. doi:10.1016/S0168-1656(01)00418-7

Xu, C., Liu, L., Zhang, Z., Jin, D., Qiu, J., and Chen, M. (2013). Genome-scale metabolic model in guiding metabolic engineering of microbial improvement. *Appl. Microbiol. Biotechnol.* 97, 519–539. doi:10.1007/s00253-012-4543-9

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.

Received: 12 September 2014; accepted: 30 December 2014; published online: 26 January 2015.

Citation: Hartmann A and Schreiber F (2015) Integrative analysis of metabolic models – from structure to dynamics. Front. Bioeng. Biotechnol. ***2****:91. doi: 10.3389/fbioe.2014.00091*

This article was submitted to Systems Biology, a section of the journal Frontiers in Bioengineering and Biotechnology.

ORIGINAL RESEARCH
published: 08 October 2015
doi: 10.3389/fbioe.2015.00154

Modeling the contribution of allosteric regulation for flux control in the central carbon metabolism of *E. coli*

Daniel Machado[1]*, *Markus J. Herrgård*[2] *and Isabel Rocha*[1]

[1] Centre of Biological Engineering, University of Minho, Braga, Portugal, [2] The Novo Nordisk Foundation Center for Biosustainability, Technical University of Denmark, Hørsholm, Denmark

Modeling cellular metabolism is fundamental for many biotechnological applications, including drug discovery and rational cell factory design. Central carbon metabolism (CCM) is particularly important as it provides the energy and precursors for other biological processes. However, the complex regulation of CCM pathways has still not been fully unraveled and recent studies have shown that CCM is mostly regulated at post-transcriptional levels. In order to better understand the role of allosteric regulation in controlling the metabolic phenotype, we expand the reconstruction of CCM in *Escherichia coli* with allosteric interactions obtained from relevant databases. This model is used to integrate multi-*omics* datasets and analyze the coordinated changes in enzyme, metabolite, and flux levels between multiple experimental conditions. We observe cases where allosteric interactions have a major contribution to the metabolic flux changes. Inspired by these results, we develop a constraint-based method (arFBA) for simulation of metabolic flux distributions that accounts for allosteric interactions. This method can be used for systematic prediction of potential allosteric regulation under the given experimental conditions based on experimental data. We show that arFBA allows predicting coordinated flux changes that would not be predicted without considering allosteric regulation. The results reveal the importance of key regulatory metabolites, such as *fructose-1,6-bisphosphate*, in controlling the metabolic flux. Accounting for allosteric interactions in metabolic reconstructions reveals a hidden topology in metabolic networks, improving our understanding of cellular metabolism and fostering the development of novel simulation methods that account for this type of regulation.

Keywords: metabolism, systems biology, constraint-based modeling, allosteric regulation, *Escherichia coli*

OPEN ACCESS

Edited by:
Firas H. Kobeissy,
University of Florida, USA

Reviewed by:
Osbaldo Resendis-Antonio,
Instituto Nacional de Medicina Genomica (INMEGEN), Mexico
Jeffrey Varner,
Purdue University, USA

Correspondence:
Daniel Machado,
Centre of Biological Engineering,
University of Minho,
Braga 4710-057, Portugal
dmachado@deb.uminho.pt

Specialty section:
This article was submitted to Systems Biology, a section of the journal Frontiers in Bioengineering and Biotechnology

Received: *16 July 2015*
Accepted: *22 September 2015*
Published: *08 October 2015*

Citation:
Machado D, Herrgård MJ and Rocha I (2015) Modeling the contribution of allosteric regulation for flux control in the central carbon metabolism of E. coli.
Front. Bioeng. Biotechnol. 3:154.
doi: 10.3389/fbioe.2015.00154

1. Introduction

Mathematical models of metabolism have become a fundamental tool for understanding cellular behavior and for designing genetic or environmental modifications to change that behavior toward a specific purpose (Heinemann and Sauer, 2010). Metabolic models have found applications in both biomedical research and industrial biotechnology. Examples of applications in biomedicine include using metabolic models of human cells to analyze the altered behavior of cancer cells and to suggest potential drug targets (Folger et al., 2011). In the context of industrial biotechnology, models of microbial metabolism are widely used for rational design of microbial cell factories (Zomorrodi et al., 2012).

There are two major approaches for modeling cellular metabolism, namely, kinetic modeling and constraint-based modeling (Machado et al., 2012). The former, based on kinetic rate laws, requires extensive experimental data for determination of the enzymatic mechanisms and respective kinetic parameters. For that reason, these models have been limited to central pathways of well-studied organisms, such as *Escherichia coli* and *Saccharomyces cerevisiae* (Teusink et al., 2000; Chassagnole et al., 2002). Constraint-based modeling, on the other hand, only accounts for the stoichiometry and directionality of biochemical reactions, which can be obtained from genome annotations and limited other information for the organism (Bordbar et al., 2014). With the increasing number of fully sequenced genomes for multiple organisms, the number of genome-scale metabolic reconstructions suitable for constraint-based modeling is also rapidly increasing, with over a hundred reconstructions currently available (Monk et al., 2014).

Constraint-based models can be used to estimate the steady-state flux distribution of a metabolic network, using the so-called Flux Balance Analysis (FBA) approach (Orth et al., 2010). Since the flux solution is not unique with only stoichiometric and directionality constraints, in FBA a single solution is selected based on the assumption of an evolutionary principle of optimality, such as maximization of cellular growth. Methods have been developed to refine metabolic flux predictions by integration of metabolic models with models of other biological processes, such as signaling and transcriptional regulatory networks (Gonçalves et al., 2013). However, some limitations of these methods, such as the reduction of gene expression levels to Boolean states, hamper the predictive ability of the integrated models. More recently, several approaches were developed to directly integrate gene expression data into metabolic models. These methods are based on the assumption that reaction fluxes should be proportional to their respective gene expression levels. However, a recent systematic evaluation of these methods showed little improvement in simulation accuracy when gene or protein expression data are used for flux prediction with a wide range of proposed methods (Machado and Herrgård, 2014). One of the conclusions from this study is that the assumption of proportionality between gene expression levels and reaction rates is not valid for many reactions.

The conclusion that transcriptional or translational regulation does not significantly regulate metabolic fluxes is consistent with recent experimental observations in multiple organisms showing that central carbon metabolism is mostly regulated at post-transcriptional levels (Daran-Lapujade et al., 2007; Chubukov et al., 2013; Kochanowski et al., 2013a). Regulation analysis is a method introduced by ter Kuile and Westerhoff (2001) for quantitatively decomposing flux regulation into *hierarchical* and metabolic coefficients. The former accounts for transcriptional and translational regulation as well as post-translational modifications, whereas the latter accounts for allosteric regulation and thermodynamics. The application of this method to three parasitic protists showed that regulation of glycolytic fluxes is never completely hierarchical, being mostly metabolic in many cases. Similar conclusions were obtained by applying this method to *S. cerevisiae*, where it was observed that metabolic regulation contributed to 50–80% of the flux change in glycolytic enzymes for the given cultivation conditions (Daran-Lapujade et al., 2007).

The partial contribution of transcriptional regulation for flux control in central carbon metabolism can be explained by the cellular trade-off between lowering the investment of protein synthesis (keeping enzymes saturated), and the need to achieve fast regulatory responses and maintain metabolic homeostasis under environmental changes (Fendt et al., 2010; Wessely et al., 2011). In fact, metabolite measurements in *E. coli* and *S. cerevisiae* have shown that most enzymes in central carbon metabolism are not saturated, with substrate levels being close to their respective K_M values (Bennett et al., 2009; Fendt et al., 2010). A recent study in *B. subtilis* showed that transcriptional regulation is insufficient to explain the observed flux change for growth in different carbon sources (Chubukov et al., 2013). Interestingly, the authors observed that the changes in substrate concentrations were also insufficient to explain the observed flux change, leaving an important contribution for post-translational modifications and allosteric regulation.

Learning how allosteric regulation controls the metabolic flux is fundamental for understanding cellular metabolism. Given the growing scope of the constraint-based modeling approach, we propose to expand this formalism with an explicit representation for allosteric interactions. In this work, we build a constraint-based model of allosteric regulation in the central carbon metabolism of *E. coli* and use it to analyze the role of this type of regulation for controlling the metabolic flux under different perturbations.

Allosteric information data are collected from relevant databases and used to build a constraint-based model expanded with allosteric interactions. We analyze how this new layer of interactions affects the network topology in terms of node connectivity and identify relevant metabolic hubs. The model is used as a scaffold to perform regulation analysis using multiple omics data for *E. coli*. Finally, a new method for constraint-based simulation accounting for allosteric interactions is proposed and used for model-based prediction of regulatory effects on flux control.

2. Results

2.1. Model Reconstruction

In order to analyze the effects of allosteric regulation in the central carbon metabolism, we expanded a constraint-based model of the core metabolism of *E. coli* (Orth et al., 2009) with allosteric interactions obtained from relevant sources (see Figure S3 in Supplementary Material and Methods section for details). The expanded model is presented in **Figure 1**. It can be observed that the integration of regulatory interactions reveals an intricate topology that is not captured by the stoichiometric reconstruction alone. In this case, the connections represent signal flow rather than mass flow. Much like in the case of signaling pathways, it is possible to observe a highly complex *crosstalk* between different subpathways. This includes multiple feedback links between upper and lower glycolysis, upper glycolysis and the oxidative part of the pentose-phosphate (PP) pathway, lower glycolysis and the TCA cycle, and a positive feedback link from citrate to upper glycolysis.

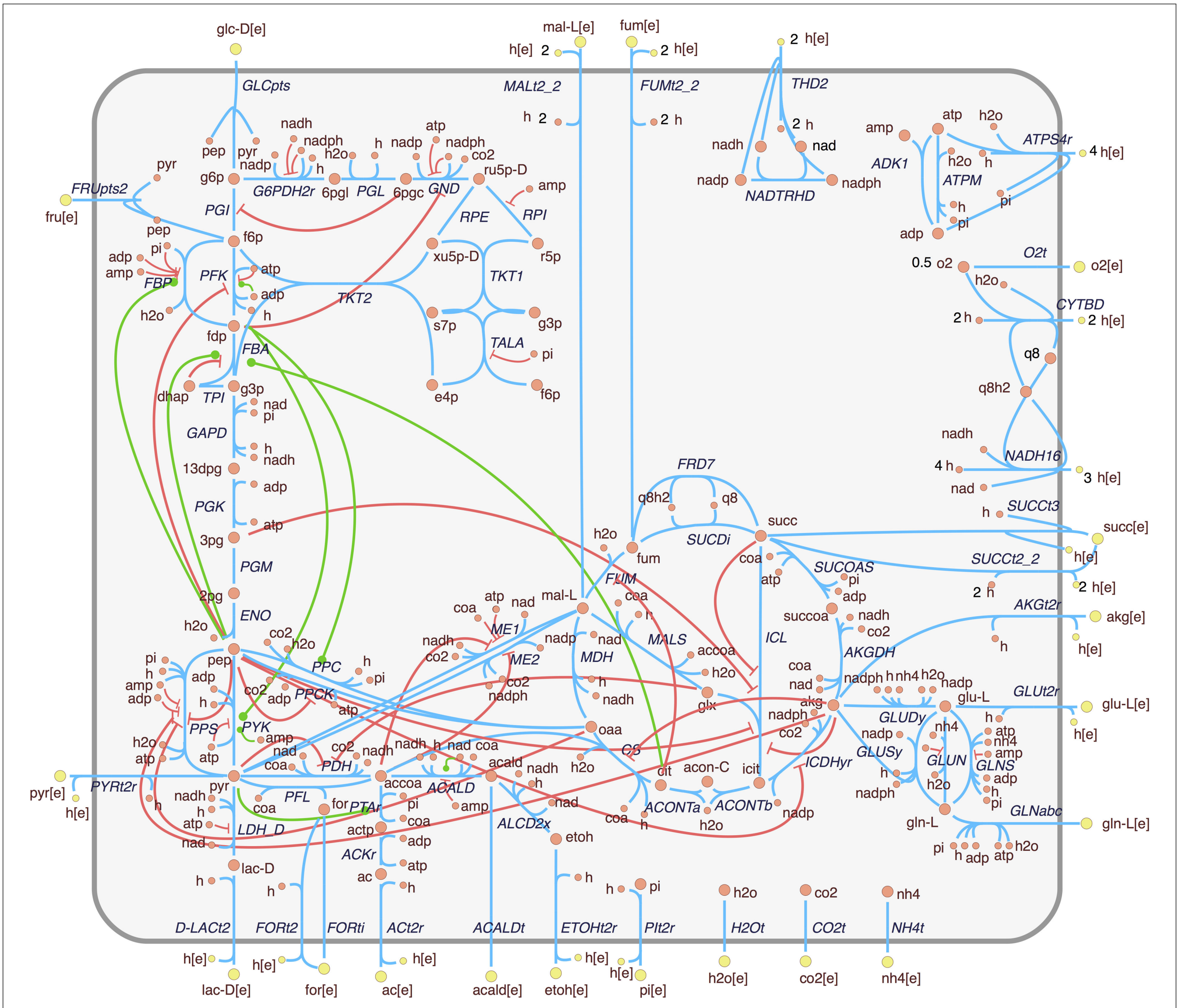


FIGURE 1 | Extension of the *E. coli* core metabolism model with allosteric interactions. Enzyme activations and inhibitions are represented, respectively, by green edges with circle ends and red edges with bar ends. This figure is adapted from the metabolic map available at the BiGG database (Schellenberger et al., 2010).

Figure 1 shows that most regulatory interactions are inhibitory. It is possible that some of these inhibitory interactions are competitive rather than allosteric (i.e., the binding site of the effector coincides with the catalytic site). Since the binding mechanisms are not generally reported in the databases, and the regulatory effect is similar, this distinction will be disregarded for the purpose of this work.

Topological analysis in terms of connectivity degree shows an increased connectivity for several metabolites when allosteric regulation is considered (**Figure 2**). However, the median value of connectivity remains the same (4 connections per metabolite). Unsurprisingly, there is an increased connectivity for metabolites that were previously known metabolic hubs. For instance, phosphoenolpyruvate (*pep*) is now connected to a total of 13 reactions (previously 8), reinforcing the importance of this glycolytic compound as a metabolic hub (Link et al., 2013; Matsuoka and Shimizu, 2015). However, changes are also observed for lowly connected metabolites. A notable case is *fructose-1,6-bisphosphate* (*fdp*), which can now be considered as a hub metabolite (with a total of 6 connections), although its connectivity is bellow the median if regulation is not considered. This metabolite was recently identified as a key flux-signaling metabolite in the glycolytic flux-sensing mechanism of *E. coli* (Kochanowski et al., 2013b).

2.2. Omics Data-Based Analysis of Allosteric Regulation

In order to understand how the coordination between hierarchical and metabolic regulation drives the metabolic flux, we used the reconstructed model to integrate and analyze a multi-*omics* dataset for *E. coli* (Ishii et al., 2007). This dataset contains transcript, protein, metabolite, and flux data for *E. coli* strains

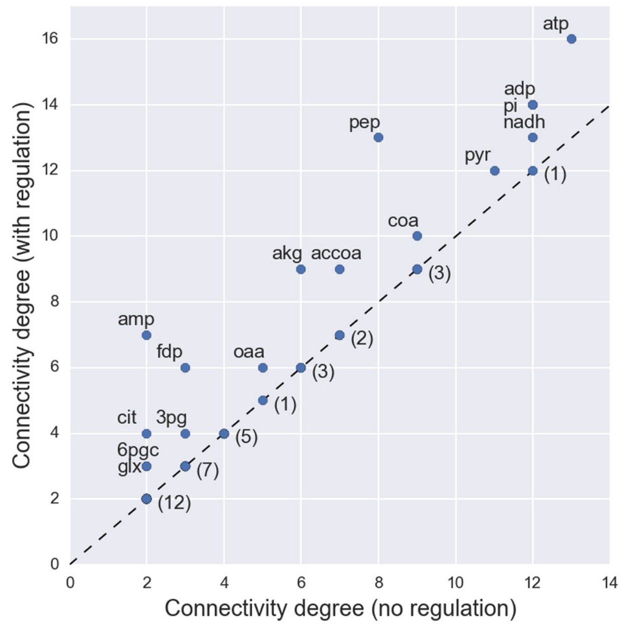


FIGURE 2 | Changes in the connectivity degree of each metabolite when allosteric regulation is considered in the network topology. The labeled metabolites represent the cases where the metabolite acts as regulator to a set of enzymes and, consequently, an increase in connectivity is observed. For the nodes with unchanged connectivity only the number of occurrences is presented.

growing aerobically in a chemostat. It comprises several experiments, including variations of dilution rate for the wild-type strain (0.1–0.7 h^{-1}) and 24 single knockout mutants growing at the reference dilution rate (0.2 h^{-1}). Herein, we will refer to the wild-type strain growing at 0.2 h^{-1} as the reference condition, and the remaining as the perturbed conditions (28 in total).

The data were analyzed using the concept of *regulation analysis* introduced by ter Kuile and Westerhoff (2001) to decompose the contribution of hierarchical (ρ_h) and metabolic (ρ_m) control coefficients during flux change between two experimental conditions ($\rho_h + \rho_m = 1$). We applied the generalization proposed in Chubukov et al. (2013) to simultaneously compare multiple conditions (see Methods). This generalization assumes that the coefficients are conserved across conditions. The results are presented in Figure S4 in Supplementary Material. It can be observed that in many cases the slopes are close to zero or even negative, indicating poor evidence of transcriptional control. Only three reactions (*PGI*, *CS*, *FUM*) present an estimated hierarchical control coefficient above 0.5. Hence, only these reactions are likely to be predominantly regulated at the transcriptional level.

Given the lack of evident hierarchical control for most enzymes, one can try to analyze the allosteric control exerted by single effectors in a similar fashion (see Methods). The results are presented in Figure S5 in Supplementary Material. In order to observe active flux control, positive slopes would be expected for enzyme activators and negative slopes for enzyme inhibitors. However, this behavior can only be observed in a few cases. The flux of *FBA* positively correlates with its two activators, citrate and *pep*. Some correlation is also observed between ATP levels and two of its inhibition targets, *GND* and *PFK*.

Given the large number of reactions without evident transcriptional or allosteric control, we hypothesize that the assumption of constant control coefficients across all conditions does not hold for the given experimental conditions. It is likely that, during different perturbations, different kinds of control are predominant for each reaction. This has also been observed in previous studies in *S. cerevisiae* (Rossell et al., 2006).

We analyzed the flux change for each reaction at each perturbed condition individually, by comparing the logarithmic change of enzyme, flux, and metabolite levels between all 28 perturbed conditions and the reference condition. Although this would result in a total of 672 potential case studies (24 regulated reactions times 28 perturbations), due to the sparsity of the data (especially the metabolome data), this study was restricted to all reaction-condition pairs with sufficient data to perform a meaningful analysis (see Methods). This reduced the number of case studies to 38 (see Figure S6 in Supplementary Material for details). We then analyzed the evidence of allosteric control for these cases (see Methods) and observed a total of 8 cases where allosteric regulation seems to play a role in controlling the reaction flux for the given perturbation (Figure S6 in Supplementary Material). These 8 cases will be analyzed in detail below.

The regulation mechanisms of the three reactions involved (*PFK*, *PPC*, and *PYK*) are depicted in **Figure 3A**. The intricate regulation of these enzymes is evident, in particular for *PFK* and *PYK*, which are catalyzed by multiple isozymes and regulated by multiple effectors. The logarithmic change of flux and all measured intervening molecules for the selected reaction-condition pairs is presented (**Figure 3B**). It can be observed that, in most cases, the change in enzyme concentration is in the opposite direction of the flux change. For *PFK*, only one of the isozymes is measured. In the case of *PYK*, where both isozymes are measured, it can be observed that the level of one isozyme increases while the other decreases. In the few cases where the flux change follows the direction of the enzyme level, the magnitude of enzyme change is still insufficient to explain the flux change (since the reaction rate would be directly proportional to the enzyme concentration). Regarding the change in substrate levels, it can be observed that, in most cases, it is also opposite to the direction of the flux change.

The effect of allosteric control is evident in some scenarios. For instance, in the Δ*ppsA* mutant, the flux of *PPC* largely increases, despite the decrease of its only enzyme (*ppc*) and its main substrate (*pep*). This increase can be explained by the increased concentration of its allosteric activator (*fdp*). There are cases where the different allosteric regulators have a cooperative effect in flux control (e.g., *PYK* at 0.7 h^{-1}) and cases where there is a competing effect (e.g., *PFK* at 0.4 h^{-1}). One can observe that flux change is not always controlled by the same combination of effectors. For instance, at high dilution rates (0.4–0.5 h^{-1}) the flux of PFK increases with the decrease of its inhibitors (ATP and *pep*), despite the decrease of its activator (ADP). However, at an even higher dilution rate (0.7 h^{-1}), the flux increase coincides with higher levels of the activator, whereas the two inhibitors change in opposite directions.

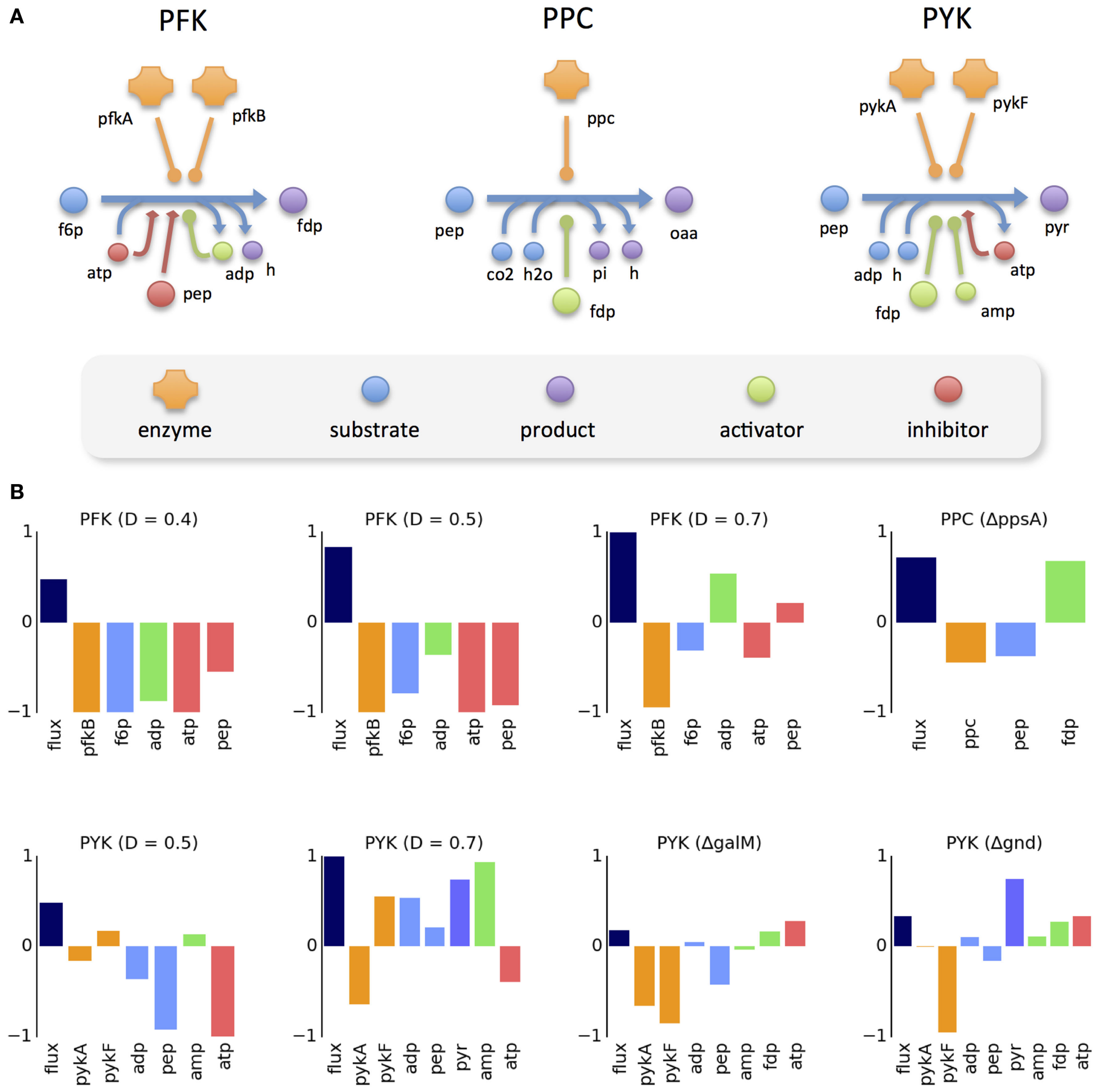


FIGURE 3 | Data analysis of allosterically regulated reactions. (A) Known regulation mechanism of three reactions analyzed in detail, including all participating molecules. **(B)** Logarithmic change of the metabolic flux and concentrations of the participating molecules between the perturbed and reference condition. Missing proteins and metabolites in the plots correspond to cases where the data were not available.

The interpretation of the results is hampered by the lack of protein and metabolite measurements for many experimental conditions. One cannot exclude the possibility that some flux changes are also driven by changes in unmeasured isozymes, cofactors, or reaction products.

2.3. Model-Based Prediction of Allosteric Regulation

Given the scarcity of multi-*omics* datasets with all the data required to perform a quantitative analysis of allosteric regulation, we developed a constraint-based approach for model-based predictions. This method is based on the assumption that, if a reaction is activated (respectively, inhibited) by a compound present in a pathway, then its flux change should be positively (respectively, negatively) correlated with the flux change in that pathway (see Supplementary Material for details). It has been proposed that allosteric intermediates function as flux-signaling metabolites that directly translate flux information to metabolite concentration (Kotte et al., 2010; Matsuoka and Shimizu, 2015). The method, named allosteric regulation FBA (arFBA), is a variation of parsimonious FBA (pFBA) (Lewis et al., 2010) where the objective function is extended as follows:

$$\min_{v} \sum_{i} |v_i| + \sum_{R_{ij}>0} w_{ij} \left| \frac{v_j}{v_j^0} - \frac{t_i}{t_i^0} \right| + \sum_{R_{ij}<0} w_{ij} \left| \frac{v_j}{v_j^0} + \frac{t_i}{t_i^0} - 2 \right| .$$

Here, v is the flux distribution to be estimated, v^0 is the flux distribution for a given reference condition, t_i is the turnover rate of metabolite i. The allosteric interactions are represented in a new matrix R, which has a structure similar to the stoichiometric matrix, with $R_{ij} = 1$ (respectively, -1) if metabolite i activates (respectively, inhibits) reaction j, and 0 otherwise (note that the stoichiometric matrix S is not changed). The w_{ij} parameters are arbitrary weights that represent the strength of the interaction between effector i and reaction j. If all w_{ij} are close to zero, then the method defaults to a simple pFBA simulation. The minimization of the extra terms in the objective function affects the respective fluxes when regulation is active. For an activation, the subtraction forces the flux and turnover ratios to be the same. For an inhibition, the term forces that a change in the turnover is compensated by an opposite change in the flux. A detailed justification for these terms is given in the Supplementary Material. The full implementation of the method is slightly more complex due to the presence of reversible reactions and reactions without flux in the reference condition (see Supplementary Material for a complete description).

In general, it is not possible to know the strength of the allosteric interactions beforehand. Therefore, we implemented an ensemble modeling approach in order to find the most plausible models (**Figure 4**). The approach is similar, albeit different, to the ensemble modeling approach used for kinetic modeling (Tran et al., 2008). A model ensemble was built by randomly sampling the w_{ij} parameters (see Methods). The simulated flux distributions are then compared with the intracellular flux data from Ishii et al. (2007). The accuracy of each model is given by the (L_1-norm) distance between the predicted and measured flux distributions. The original ensemble is split into two groups containing the models with prediction accuracy above and below the median. We then perform enrichment analysis by comparing the distributions of each parameter between the two ensembles. For a particular experimental condition, if a parameter w_{ij} has systematically higher values in the ensemble with higher predictive accuracy, then the assumption of allosteric control between effector i and reaction j results in improved flux predictions for that condition.

Figure 5 shows t-test values for all parameters across all experimental conditions. Although there are not clearly defined clusters in the clustered heatmap, some general patterns can be observed. About one-quarter of the interactions are positively enriched for most experimental conditions, representing probable cases of active allosteric control for those conditions. On the other hand, almost half of the parameters are negatively enriched for a majority of conditions. These represent allosteric constraints that, in most cases, hamper the predictive ability of the models. Finally, there is a subset of allosteric interactions which are neither positively nor negatively enriched. Accounting for these interactions has very little effect in the prediction of flux distributions for the given experimental conditions.

The most frequent positively enriched interactions include inhibition of the oxidative phase of the pentose-phosphate pathway (PPP) by reducing agents NADH and NADPH; mutual inhibition between PPP and upper glycolysis; feedforward activation of *PPC* and *PYK* by *fdp*; and inhibition of the glyoxylate shunt by multiple effectors. Interestingly, several parameters that are positively enriched for a subset of conditions are also negatively enriched for some of the remaining conditions. Hence, although the respective interactions improve the flux predictions in some conditions, in other conditions they make predictions worse.

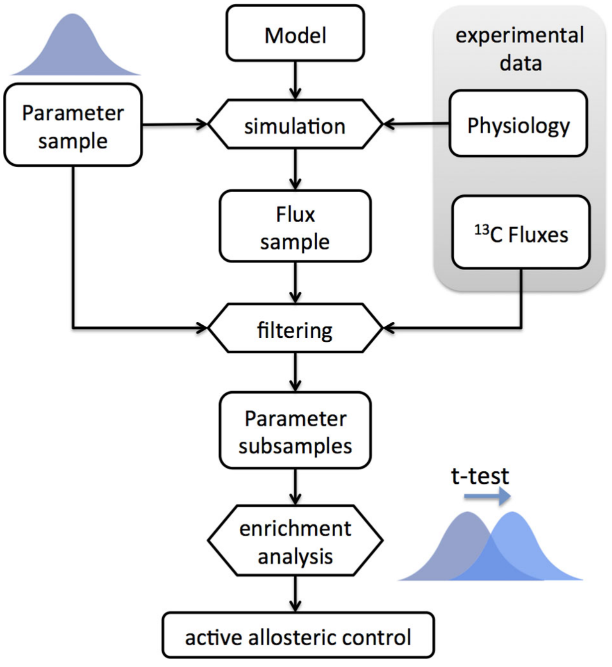


FIGURE 4 | Workflow diagram of the enrichment analysis based on the ensemble modeling method. An ensemble of models is built by random sampling of the parameter space (log-normal distribution). Physiological data (growth and uptake rates) are used to constrain the models. The ensemble is used for simulation of flux distributions, which are filtered by comparison with ^{13}C-based intracellular flux data. The subset of ensembles with higher predictive ability is compared to those with lower predictive ability and enriched parameters are detected by t-test analysis. The active allosteric control cases are identified by the positively enriched parameters for the respective interactions.

In order to test the predictive ability of our *in silico* approach, we analyzed the enrichment results for the potential cases of allosteric control previously detected by data-driven analysis (**Figure 3B**). Some of the allosteric interactions were significantly enriched, namely the activation of *PFK* by ADP at the highest dilution rate ($t = 4.28$, $p = 1.88\text{e-}5$), activation of *PPC* by *fdp* in the $\Delta ppsA$ mutant (t = 19.0, $p = 8.17\text{e-}79$), and activation of *PYK* by *fdp* in the Δgnd mutant ($t = 4.70$, $p = 2.63\text{e-}6$) and the $\Delta galM$ mutant ($t = 7.09$, $p = 1.44\text{e-}12$).

It should be noted that we are using our simulation method (arFBA) in the reverse direction, i.e., a model ensemble is compared with experimental data to find which parameters (weighting factors) result in improved predictions. Although, in theory, one could use the method in the forward direction, i.e., to perform simulations with improved flux predictions, this would require finding a "universal" parameter configuration that fits all conditions. The previous results show that such universal configuration cannot be found due to the condition-specific nature of allosteric regulation. Nonetheless, we tested the accuracy of arFBA by measuring the distance between simulated and experimental flux distributions. **Figure 6** shows the frequency distribution of the distances obtained by random sampling of the weighting factors for each experimental condition. The distance obtained with FBA is shown for comparison. It can be observed that,

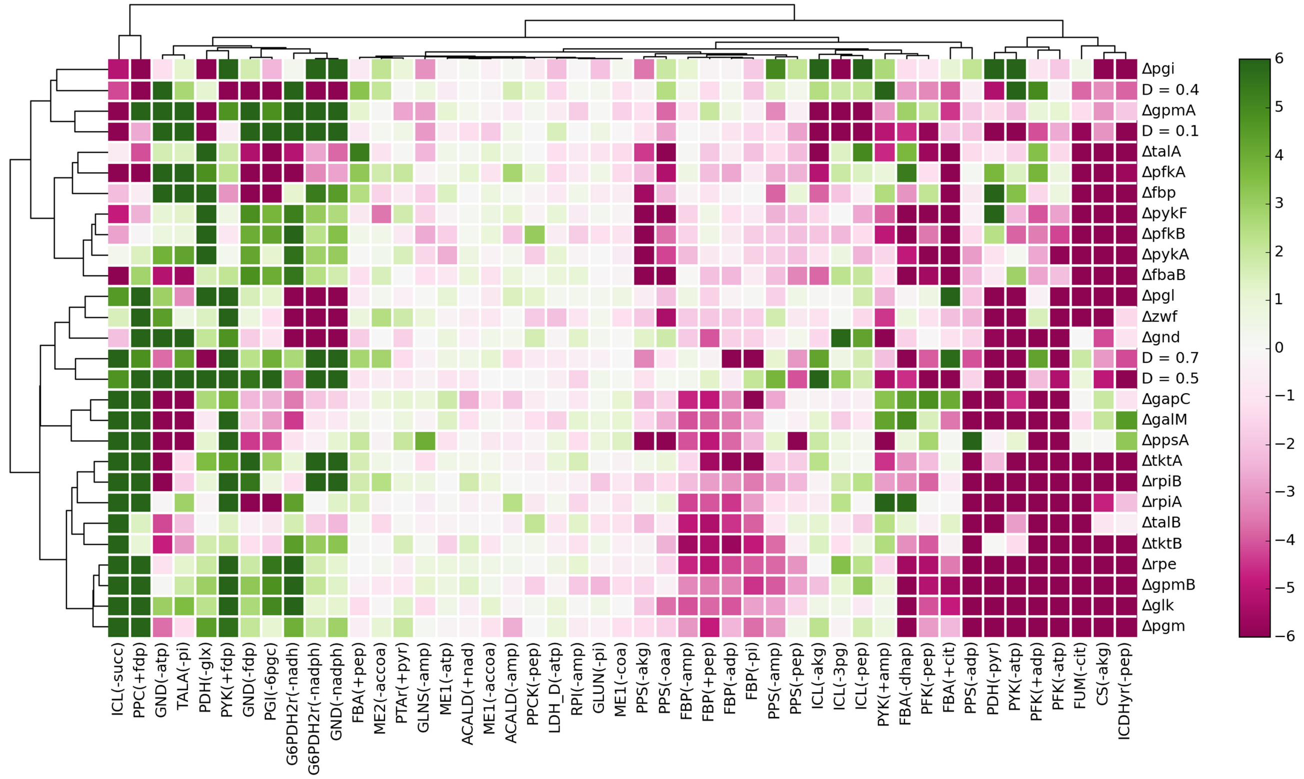


FIGURE 5 | Enrichment analysis of the parameters associated with each allosteric interaction, represented by the *t*-test value of each parameter subsample for each experimental condition. The clustering of the heatmap was performed using complete linkage and Manhattan distance.

for most experimental conditions, the average distance obtained with arFBA is lower than that obtained with FBA, indicating a higher accuracy of the former. Finally, we tested the accuracy of arFBA with *a posteriori* calibration of the weighting factors (see methods). It can be observed that, after calibration, the accuracy of arFBA is higher than FBA for 26 of the 28 conditions.

3. Discussion

In this work, we analyzed the role of allosteric regulation for flux control in the central carbon metabolism of *E. coli*. For this, we extended a constraint-based metabolic model of *E. coli* with allosteric regulation. The application of such a model is twofold. First, it can be used as an integrative scaffold for multi-*omics* dataset analysis, revealing the coordination between enzyme, metabolite, and flux levels. Second, it can be used for *in silico*-based predictions that account for allosteric regulation in the simulation of the metabolic phenotype. For that purpose, we implemented an FBA variant, named arFBA, that accounts for allosteric interactions in the determination of the flux distribution.

Using the expanded model and a multi-*omics* dataset for *E. coli* (Ishii et al., 2007), we analyzed the impact of allosteric regulation in controlling the metabolic flux under multiple environmental and genetic perturbations. We implemented a generalized form of regulation analysis (ter Kuile and Westerhoff, 2001) in order to find which reactions are predominantly under transcriptional or allosteric control. The results reveal that most reactions are generally not controlled by the same mechanism across all conditions. This led us to analyze the effects of perturbations in single reactions for each experimental condition. This analysis is hampered by missing protein and metabolite measurements, which does not allow accounting for all participating compounds in the reactions analyzed. Although we neglected the effect of missing isozyme and cofactor measurements, as well as product concentrations for irreversible reactions, only 38 out of 672 possible case studies (24 reactions × 28 perturbations) could be analyzed in a meaningful way (Figure S6 in Supplementary Material). Nonetheless, it was possible to identify 8 (out of 38) cases where the reaction flux is predominantly controlled by allosteric mechanisms.

Considering that the dataset published by Ishii et al. (2007) is one of the most comprehensive multi-*omics* dataset for a model organism published so far, we can conclude that purely data-driven analysis is very limited for studying metabolic regulation. Therefore, we applied our simulation method using an ensemble modeling approach to identify which allosteric interactions result in improved flux predictions. Enrichment analysis of the weighting factors in our model revealed that several allosteric interactions were significantly enriched when the models were filtered by their agreement with experimental flux data. A comparison between the *in silico* results and the data-driven analysis showed that 4 of the 8 cases of allosteric control previously identified were also detected by the computational approach.

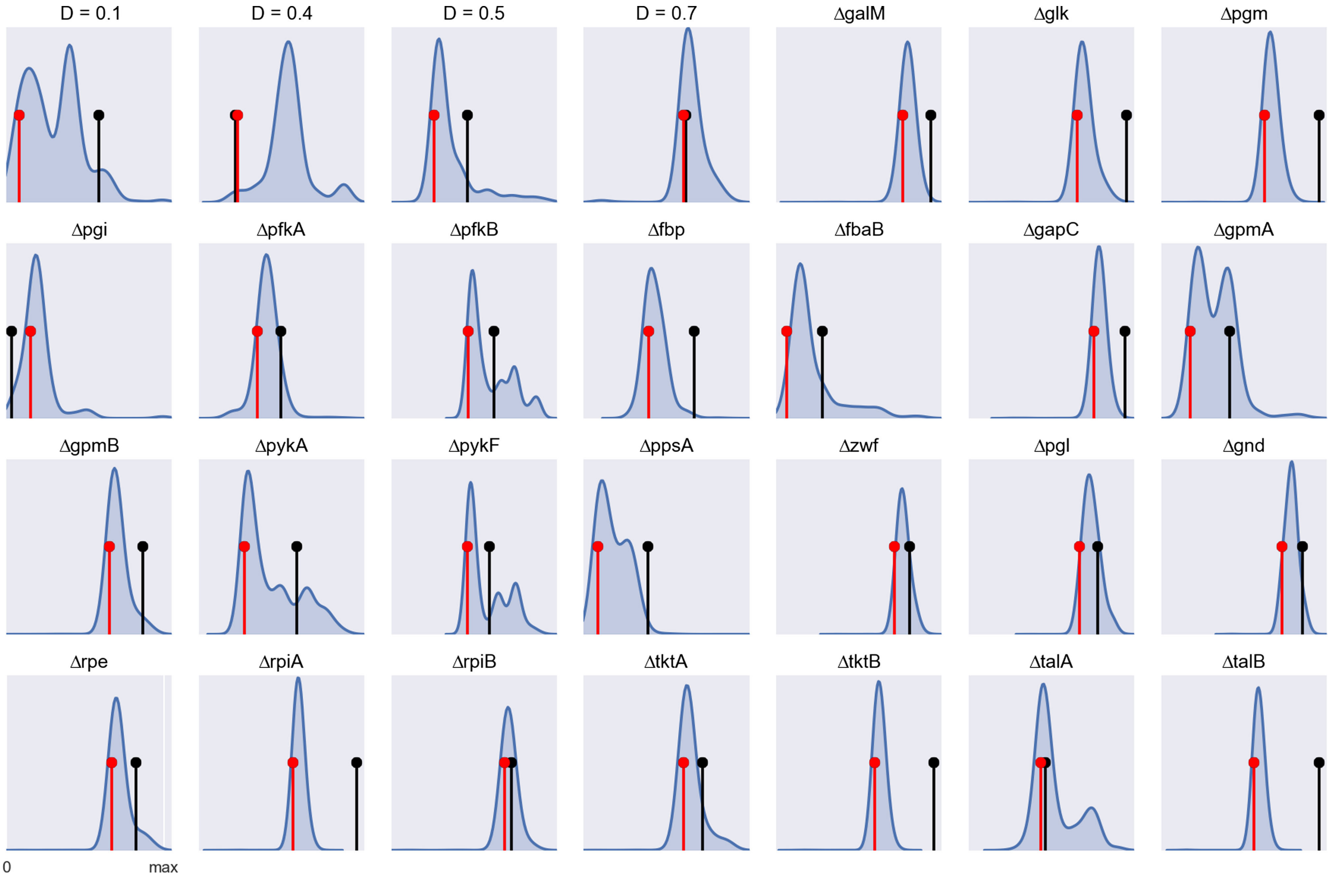


FIGURE 6 | Simulation accuracy determined by the (L_1-norm) distance between experimental and simulated flux distributions for each experimental condition (data normalized by the maximum distance value). The blue curve shows the frequency distribution of the distances obtained by random sampling of the weighting factors in arFBA. The black pin marks the distance obtained with an FBA simulation. The red pin marks the distance obtained with arFBA after calibration of the weighting factors.

Given the very limited scope of the cases analyzed in detail, the cross-comparison between the data driven and *in silico* results can hardly be considered a validation of the latter. In order to determine the accuracy of the simulation method it would be necessary to estimate the number of false positive and false negative results for the whole dataset. Instead, the two approaches should be seen as complementary methods to guide the analysis of allosteric regulation. Furthermore, the data analysis revealed that the predominant mode of regulation for each reaction is condition dependent. This was also observed in the *in silico* analysis, hampering the determination of a universal set of weighting factors for arFBA. Given the interplay between different regulation mechanisms, the approach developed herein could be suitable for integration with other methods for identification of regulation mechanisms (Bordel et al., 2010).

An ensemble modeling approach was also employed by Link et al. (2013) for systematic identification of allosteric interactions in *E. coli*. The authors measured metabolite concentrations using rapid sampling and ^{13}C-labeled substrates (glucose and fructose) to determine the transient profile of glycolytic intermediates in dynamic cultures switching between glycolysis and gluconeogenesis. A kinetic ensemble model for glycolysis was used to test 126 putative interactions. The results not only confirmed previously known interactions but also predicted new interactions that had not been previously reported. Although the model used in this study differs from ours, the results regarding interactions common to both models are consistent. In particular, both studies revealed the importance of *PFK* as an active regulation target for controlling the glycolytic flux, and the role of *fdp* as key regulator of *PPC* and *PYK* to control *pep* consumption.

At the end of our data-driven analysis, some flux changes remain unexplained by hierarchical or metabolic control. One main reason for this is the lack of coverage of the metabolomics data, which only accounts for approximately half of the metabolites in the model. Another possibility is that the regulatory mechanisms for the respective enzymes are not fully known or the relevant allosteric interactions were not included in the model. It is also possible that the enzyme concentrations do not correlate with the respective enzymatic activity due to post-translational modifications (PTMs). It has been shown that PTMs, such as acetylation, have important regulatory functions in *E. coli* (Castaño-Cerezo et al., 2014).

The generation of high-quality multi-*omics* datasets will be necessary for a deeper understanding of metabolic regulation.

Herein, we used a previously published dataset for chemostat cultures. However, steady-state data may be insufficient to analyze regulatory responses. It has been observed that fast metabolic responses precede the slower transcriptional response during metabolic adaptation (Ralser et al., 2009). Since allosteric regulation operates on a faster time-scale compared to transcriptional regulation, transient profiles on short time scales should be particularly informative (Link et al., 2013).

4. Conclusion

In this work, we focused on the role of allosteric regulation in central carbon metabolism. The reconstruction of an allosteric model revealed that allosteric information is inconsistent among different data sources even for these highly studied pathways. The allosteric interactions added a new layer to the network topology, changing the overall network connectivity and revealing metabolic hubs that would otherwise be ignored (e.g., *fdp*). Hierarchical and allosteric regulation analysis using a multi-*omics* dataset revealed that there is no predominant mechanism of regulation across all experimental conditions. Nonetheless, situations of predominant allosteric control could be identified for some reactions at particular conditions. Our new method for model-based prediction of allosteric control was able to capture at least a few of these situations. However, the assessment of the predictive ability of this method is hampered by the lack of more comprehensive data.

For central carbon metabolism, it would have been feasible to perform this analysis using a kinetic modeling approach [similarly to Link et al. (2013)]. However, as we move toward regulatory analysis at the genome-scale, the constraint-based approach should become especially useful. Building a genome-scale model of allosteric regulation is a daunting task that will require literature mining, extensive manual curation, and prediction of putative interactions. Our knowledge of the *allosterome* is currently limited by the lack of high-throughput screening methods for detecting metabolite–enzyme interactions. It is likely that the vast majority of allosteric interactions are yet to be discovered (Lindsley and Rutter, 2006). Recent experimental methods have been developed toward systematic identification of metabolite-protein interactions (Gallego et al., 2010; Li et al., 2010; Orsak et al., 2011; Feng et al., 2014). However, we are still far from a genome-scale screening of the hundreds of thousands of potential interactions between all metabolites and enzymes in an organism.

Notebaart et al. (2014) have recently unraveled the *underground* metabolism of *E. coli* by expanding a genome-scale metabolic model with reactions resulting from promiscuous enzyme activity. With the *allosterome*, we can unravel yet another hidden layer in the network topology of cellular metabolism. New expanded models of metabolism will be certainly useful for applications, such as drug discovery and rational strain design, as we slowly move toward what has been called the "second secret of life" (Fenton, 2008).

A python implementation of arFBA as well as the allosteric model in SBML format are available on GitHub: https://github.com/cdanielmachado/arfba.

5. Materials and Methods

5.1. Model Reconstruction

The original model of the core metabolism of *E. coli* (Orth et al., 2009) was extended with allosteric interactions obtained from BRENDA (Schomburg et al., 2002), EcoCyc (Keseler et al., 2011), and two previously published kinetic models (Chassagnole et al., 2002; Kotte et al., 2010). We searched for evidence of regulatory interactions for each possible combination of enzymes and metabolites in the model. A total of 148 regulatory interactions were found (Figure S3 in Supplementary Material). Since the majority of these interactions can only be found in one data source, for the sake of curation we only included in the model the interactions that are reported in at least two different sources. In a few cases the same metabolite is reported as activator and inhibitor of an enzyme (e.g., *phosphoenolpyruvate* binding to *fructose-bisphosphatase*). In these cases, we used the most frequently reported effect.

5.2. Regulation Analysis

5.2.1. Cross-Condition Analysis

The metabolic flux of a reaction (J_i) can be generically described in terms of the concentrations of the respective enzyme(s) (E_i) and all the intervening metabolites (substrates, products, effectors):

$$J_i = k_{\text{cat}} E_i f(M)$$

where k_{cat} is the turnover rate of the enzyme, and $f(M)$ represents a non-linear function of the metabolite concentrations. *Regulation analysis* introduced by ter Kuile and Westerhoff (2001) decomposes the contribution from hierarchical and metabolic control by considering the logarithmic change between two experimental conditions:

$$\Delta \log(J_i) = \Delta \log(E_i) + \Delta \log(f(M))$$

and estimating the respective contribution coefficients:

$$1 = \frac{\Delta \log(E_i)}{\Delta \log(J_i)} + \frac{\Delta \log(f(M))}{\Delta \log(J_i)} = \rho_h + \rho_m.$$

Since $f(M)$ is generally unknown, one can estimate ρ_h (and consequently ρ_m) by measuring the enzyme and flux levels across different conditions. Chubukov et al. (2013) generalized this comparison from two to multiple conditions in order to decrease the effects of experimental error. The estimation is performed by linear regression between $\log(E_i)$ and $\log(J_i)$ across all experimental conditions using a robust linear regression method (Theil–Sen estimator).

We further generalized this concept to the study of allosteric regulation, by decoupling the effect of allosteric regulators in the reaction flux from the non-linear $f(M)$ component, using a power-law approximation:

$$f(M) \approx g(S, P) \prod_j A_j^{\gamma_{ij}} \prod_j I_j^{-\gamma_{ij}}$$

where S, P, A, I represent, respectively, the set of substrates, products, activators and inhibitors of reaction i, and γ_{ij} is the apparent kinetic order of effector j in reaction i, as defined in Biochemical-Systems Theory (Voit, 2013). This allows us to estimate individual allosteric regulation coefficients (ρ_a) for each effector as:

$$\rho_{\mathrm{a}(j)} = \begin{cases} \gamma_{\mathrm{ij}} \dfrac{\Delta \log(A_j)}{\Delta \log(J_i)} & \text{if } j \text{ is an activator of } i \\ -\gamma_{\mathrm{ij}} \dfrac{\Delta \log(I_j)}{\Delta \log(J_i)} & \text{if } j \text{ is an inhibitor of } i \end{cases}$$

With the exception of effectors exhibiting cooperative binding, we can assume that the kinetic orders are close to or below unity ($\gamma_{ij} \leq 1$). Hence, the allosteric control coefficient is bound by the slope of the linear regression.

Regulation analysis was performed for all allosterically regulated reactions with available fluxomics and proteomics data. A total of 18 (out of 24) regulated reactions were experimentally measured. Due to gaps in the proteomics dataset, we restricted the analysis to enzymes with available data for at least 10 (out of 29) experimental conditions.

5.2.2. Single-Condition Analysis

Allosteric effects were analyzed for each perturbation individually by comparing the logarithmic change of enzyme, flux, and metabolite levels between all 28 perturbed conditions and the reference condition. Due to the sparsity of the data (especially the metabolome data), this analysis was restricted to all reaction-condition combinations where the following criteria were satisfied: (1) at least one associated enzyme was measured; (2) all main substrates (excluding cofactors) were measured; (3) at least one effector was measured. Furthermore, we excluded flux changes that were not significant (i.e., the perturbed flux falls within a 95% confidence interval of the reference flux).

Evidence of allosteric control was detected by selecting conditions where the flux change is not fully explained by changes in enzyme concentration ($\Delta\log(E)/\Delta\log(J) < 0.5$) or substrate abundance ($\Delta\log(S)/\Delta\log(J) < 0.5$), and is at least partly related with changes in one allosteric activator ($\Delta\log(A)/\Delta\log(J) > 0.25$) or inhibitor ($-\Delta\log(I)/\Delta\log(J) > 0.25$). For reversible reactions, the effect of flux changes arising from changes in the thermodynamic driving force cannot be excluded. Therefore, for these reactions we only considered reactions where the products were experimentally measured (excluding cofactors) and the flux change cannot be fully explained by the change in product abundance ($-\Delta\log(P)/\Delta\log(J) < 0.5$).

5.2.3. Ensemble Modeling with arFBA

For each experimental condition, an ensemble of 10^4 models was built by sampling the weighting factors (w_{ij} parameters) from a log-normal distribution. Each model is constrained with the experimentally measured glucose and oxygen uptake rates, and the growth rate, which is given by the dilution rate.

5.2.4. Calibration of Weighting Factors in arFBA

Condition-specific weighting factors were calibrated for each experimental condition as follows: an ensemble of 10^4 arFBA models was built as described above; the accuracy of each model was determined by the L_1-norm distance between the experimental and simulated flux distributions; the calibrated weighting factors were calculated as the average of the 10% most accurate models.

Acknowledgments

DM and IR would like to thank the FCT Strategic Project PEst-OE/EQB/LA0023/2013 and the Project "BioInd – Biotechnology and Bioengineering for improved Industrial and Agro-Food processes", REF. NORTE-07-0124-FEDER-000028, co-funded by the Programa Operacional Regional do Norte (ON.2 – O Novo Norte), QREN, FEDER. MJH would like to thank the Novo Nordisk Foundation for support.

Supplementary Material

The Supplementary Material for this article can be found online at http://journal.frontiersin.org/article/10.3389/fbioe.2015.00154

References

Bennett, B. D., Kimball, E. H., Gao, M., Osterhout, R., Van Dien, S. J., and Rabinowitz, J. D. (2009). Absolute metabolite concentrations and implied enzyme active site occupancy in *Escherichia coli*. *Nat. Chem. Biol.* 5, 593–599. doi:10.1038/nchembio.186

Bordbar, A., Monk, J. M., King, Z. A., and Palsson, B. O. (2014). Constraint-based models predict metabolic and associated cellular functions. *Nat. Rev. Genet.* 15, 107–120. doi:10.1038/nrg3643

Bordel, S., Agren, R., and Nielsen, J. (2010). Sampling the solution space in genome-scale metabolic networks reveals transcriptional regulation in key enzymes. *PLoS Comput. Biol.* 6:e1000859. doi:10.1371/journal.pcbi.1000859

Castaño-Cerezo, S., Bernal, V., Post, H., Fuhrer, T., Cappadona, S., Sánchez-Díaz, N. C., et al. (2014). Protein acetylation affects acetate metabolism, motility and acid stress response in *Escherichia coli*. *Mol. Syst. Biol.* 10, 762. doi:10.15252/msb.20145227

Chassagnole, C., Noisommit-Rizzi, N., Schmid, J. W., Mauch, K., and Reuss, M. (2002). Dynamic modeling of the central carbon metabolism of *Escherichia coli*. *Biotechnol. Bioeng.* 79, 53–73. doi:10.1002/bit.10288

Chubukov, V., Uhr, M., Le Chat, L., Kleijn, R. J., Jules, M., Link, H., et al. (2013). Transcriptional regulation is insufficient to explain substrate-induced flux changes in *Bacillus subtilis*. *Mol. Syst. Biol.* 9, 709. doi:10.1038/msb.2013.66

Daran-Lapujade, P., Rossell, S., van Gulik, W. M., Luttik, M. A., de Groot, M. J., Slijper, M., et al. (2007). The fluxes through glycolytic enzymes in *Saccharomyces cerevisiae* are predominantly regulated at posttranscriptional levels. *Proc. Natl. Acad. Sci. U.S.A.* 104, 15753–15758. doi:10.1073/pnas.0707476104

Fendt, S.-M., Buescher, J. M., Rudroff, F., Picotti, P., Zamboni, N., and Sauer, U. (2010). Tradeoff between enzyme and metabolite efficiency maintains metabolic homeostasis upon perturbations in enzyme capacity. *Mol. Syst. Biol.* 6, 356. doi:10.1038/msb.2010.11

Feng, Y., De Franceschi, G., Kahraman, A., Soste, M., Melnik, A., Boersema, P. J., et al. (2014). Global analysis of protein structural changes in complex proteomes. *Nat. Biotechnol.* 32, 1036–1044. doi:10.1038/nbt.2999

Fenton, A. W. (2008). Allostery: an illustrated definition for the 'second secret of life'. *Trends Biochem. Sci.* 33, 420–425. doi:10.1016/j.tibs.2008.05.009

Folger, O., Jerby, L., Frezza, C., Gottlieb, E., Ruppin, E., and Shlomi, T. (2011). Predicting selective drug targets in cancer through metabolic networks. *Mol. Syst. Biol.* 7, 501. doi:10.1038/msb.2011.35

Gallego, O., Betts, M. J., Gvozdenovic-Jeremic, J., Maeda, K., Matetzki, C., Aguilar-Gurrieri, C., et al. (2010). A systematic screen for protein–lipid interactions in *Saccharomyces cerevisiae*. *Mol. Syst. Biol.* 6, 430. doi:10.1038/msb.2010.87

Gonçalves, E., Bucher, J., Ryll, A., Niklas, J., Mauch, K., Klamt, S., et al. (2013). Bridging the layers: towards integration of signal transduction, regulation and metabolism into mathematical models. *Mol. Biosyst.* 9, 1576–1583. doi:10.1039/c3mb25489e

Heinemann, M., and Sauer, U. (2010). Systems biology of microbial metabolism. *Curr. Opin. Microbiol.* 13, 337–343. doi:10.1016/j.mib.2010.02.005

Ishii, N., Nakahigashi, K., Baba, T., Robert, M., Soga, T., Kanai, A., et al. (2007). Multiple high-throughput analyses monitor the response of *E. coli* to perturbations. *Science* 316, 593–597. doi:10.1126/science.1132067

Keseler, I. M., Collado-Vides, J., Santos-Zavaleta, A., Peralta-Gil, M., Gama-Castro, S., Muñiz-Rascado, L., et al. (2011). EcoCyc: a comprehensive database of *Escherichia coli* biology. *Nucleic Acids Res.* 39(Suppl. 1), D583–D590. doi:10.1093/nar/gkq1143

Kochanowski, K., Sauer, U., and Chubukov, V. (2013a). Somewhat in control – the role of transcription in regulating microbial metabolic fluxes. *Curr. Opin. Biotechnol.* 24, 987–993. doi:10.1016/j.copbio.2013.03.014

Kochanowski, K., Volkmer, B., Gerosa, L., van Rijsewijk, B. R. H., Schmidt, A., and Heinemann, M. (2013b). Functioning of a metabolic flux sensor in *Escherichia coli*. *Proc. Natl. Acad. Sci. U.S.A.* 110, 1130–1135. doi:10.1073/pnas.1202582110

Kotte, O., Zaugg, J. B., and Heinemann, M. (2010). Bacterial adaptation through distributed sensing of metabolic fluxes. *Mol. Syst. Biol.* 6, 355. doi:10.1038/msb.2010.10

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, 1–13. doi:10.1038/msb.2010.47

Li, X., Gianoulis, T. A., Yip, K. Y., Gerstein, M., and Snyder, M. (2010). Extensive in vivo metabolite-protein interactions revealed by large-scale systematic analyses. *Cell* 143, 639–650. doi:10.1016/j.cell.2010.09.048

Lindsley, J. E., and Rutter, J. (2006). Whence cometh the allosterome? *Proc. Natl. Acad. Sci. U.S.A.* 103, 10533–10535. doi:10.1073/pnas.0604452103

Link, H., Kochanowski, K., and Sauer, U. (2013). Systematic identification of allosteric protein-metabolite interactions that control enzyme activity in vivo. *Nat. Biotechnol.* 31, 357–361. doi:10.1038/nbt.2489

Machado, D., Costa, R. S., Ferreira, E. C., Rocha, I., and Tidor, B. (2012). Exploring the gap between dynamic and constraint-based models of metabolism. *Metab. Eng.* 14, 112–119. doi:10.1016/j.ymben.2012.01.003

Machado, D., and Herrgård, M. (2014). Systematic evaluation of methods for integration of transcriptomic data into constraint-based models of metabolism. *PLoS Comput. Biol.* 10:e1003580. doi:10.1371/journal.pcbi.1003580

Matsuoka, Y., and Shimizu, K. (2015). Current status and future perspectives of kinetic modeling for the cell metabolism with incorporation of the metabolic regulation mechanism. *Bioresour. Bioprocess.* 2, 1–19. doi:10.1186/s40643-014-0031-7

Monk, J., Nogales, J., and Palsson, B. O. (2014). Optimizing genome-scale network reconstructions. *Nat. Biotechnol.* 32, 447–452. doi:10.1038/nbt.2870

Notebaart, R. A., Szappanos, B., Kintses, B., Pál, F., Györkei, Á, Bogos, B., et al. (2014). Network-level architecture and the evolutionary potential of underground metabolism. *Proc. Natl. Acad. Sci. U.S.A.* 111, 11762–11767. doi:10.1073/pnas.1406102111

Orsak, T., Smith, T. L., Eckert, D., Lindsley, J. E., Borges, C. R., and Rutter, J. (2011). Revealing the allosterome: systematic identification of metabolite-protein interactions. *Biochemistry* 51, 225–232. doi:10.1021/bi201313s

Orth, J., Fleming, R., and Palsson, B. (2009). "Reconstruction and use of microbial metabolic networks: the core *Escherichia coli* metabolic model as an educational guide," in *EcoSal – Escherichia coli and Salmonella: Cellular and Molecular Biology*, eds A. Bock, I. R. Curtiss, J. Kaper, P. Karp, F. Neidhardt, T. Nystrom, et al. (Washington, DC: ASM Press), 56–99.

Orth, J. D., Thiele, I., and Palsson, B. Ø (2010). What is flux balance analysis? *Nat. Biotechnol.* 28, 245–248. doi:10.1038/nbt.1614

Ralser, M., Wamelink, M. M., Latkolik, S., Jansen, E. E., Lehrach, H., and Jakobs, C. (2009). Metabolic reconfiguration precedes transcriptional regulation in the antioxidant response. *Nat. Biotechnol.* 27, 604–605. doi:10.1038/nbt0709-604

Rossell, S., van der Weijden, C. C., Lindenbergh, A., van Tuijl, A., Francke, C., Bakker, B. M., et al. (2006). Unraveling the complexity of flux regulation: a new method demonstrated for nutrient starvation i*n Saccharomyces cerevisiae*. *Proc. Natl. Acad. Sci. U.S.A.* 103, 2166–2171. doi:10.1073/pnas.0509831103

Schellenberger, J., Park, J. O., Conrad, T. M., and Palsson, B. Ø (2010). BiGG: a biochemical genetic and genomic knowledgebase of large scale metabolic reconstructions. *BMC Bioinformatics* 11:213. doi:10.1186/1471-2105-11-213

Schomburg, I., Chang, A., and Schomburg, D. (2002). BRENDA, enzyme data and metabolic information. *Nucleic Acids Res.* 30, 47–49. doi:10.1093/nar/30.1.47

ter Kuile, B. H., and Westerhoff, H. V. (2001). Transcriptome meets metabolome: hierarchical and metabolic regulation of the glycolytic pathway. *FEBS Lett.* 500, 169–171. doi:10.1016/S0014-5793(01)02613-8

Teusink, B., Passarge, J., Reijenga, C. A., Esgalhado, E., van der Weijden, C. C., Schepper, M., et al. (2000). Can yeast glycolysis be understood in terms of in vitro kinetics of the constituent enzymes? Testing biochemistry. *Eur. J. Biochem.* 267, 5313–5329. doi:10.1046/j.1432-1327.2000.01527.x

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

Voit, E. O. (2013). Biochemical systems theory: a review. *ISRN Biomath.* 2013, 897658. doi:10.1155/2013/897658

Wessely, F., Bartl, M., Guthke, R., Li, P., Schuster, S., and Kaleta, C. (2011). Optimal regulatory strategies for metabolic pathways in *Escherichia coli* depending on protein costs. *Mol. Syst. Biol.* 7, 515. doi:10.1038/msb.2011.46

Zomorrodi, A. R., Suthers, P. F., Ranganathan, S., and Maranas, C. D. (2012). Mathematical optimization applications in metabolic networks. *Metab. Eng.* 14, 672–686. doi:10.1016/j.ymben.2012.09.005

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.

frontiers in
BIOENGINEERING AND BIOTECHNOLOGY

REVIEW ARTICLE
published: 16 February 2015
doi: 10.3389/fbioe.2015.00013

Analysis of genetic variation and potential applications in genome-scale metabolic modeling

João G. R. Cardoso[1]*, Mikael Rørdam Andersen*[2]*, Markus J. Herrgård*[1] *and Nikolaus Sonnenschein*[1]*

[1] The Novo Nordisk Foundation Center of Biosustainability, Technical University of Denmark, Hørsholm, Denmark
[2] Department of Systems Biology, Technical University of Denmark, Lyngby, Denmark

Edited by:
Natalia Polouliakh, Sony Computer Science Laboratories Inc., Japan

Reviewed by:
Guanglong Jiang, Indiana University School of Medicine, USA
Nathan Price, Institute for Systems Biology, USA

****Correspondence:***
Nikolaus Sonnenschein, The Novo Nordisk Foundation Center of Biosustainability, Technical University of Denmark, Kogle Allé 6, Hørsholm DK-2970, Denmark
e-mail: niso@biosustain.dtu.dk

Genetic variation is the motor of evolution and allows organisms to overcome the environmental challenges they encounter. It can be both beneficial and harmful in the process of engineering cell factories for the production of proteins and chemicals. Throughout the history of biotechnology, there have been efforts to exploit genetic variation in our favor to create strains with favorable phenotypes. Genetic variation can either be present in natural populations or it can be artificially created by mutagenesis and selection or adaptive laboratory evolution. On the other hand, unintended genetic variation during a long term production process may lead to significant economic losses and it is important to understand how to control this type of variation. With the emergence of next-generation sequencing technologies, genetic variation in microbial strains can now be determined on an unprecedented scale and resolution by re-sequencing thousands of strains systematically. In this article, we review challenges in the integration and analysis of large-scale re-sequencing data, present an extensive overview of bioinformatics methods for predicting the effects of genetic variants on protein function, and discuss approaches for interfacing existing bioinformatics approaches with genome-scale models of cellular processes in order to predict effects of sequence variation on cellular phenotypes.

Keywords: genetic variation, SNP, next-generation sequencing, constraint-based modeling, metabolic engineering, adaptive laboratory evolution, metabolism, high-throughput analysis

1. INTRODUCTION

Genetic engineering has been used for several decades to manipulate microorganisms in order to allow production of valuable products, including primary metabolites (e.g., amino-acids and organic acids), secondary metabolites (e.g., antibiotics), and enzymes or other recombinant proteins (Adrio and Demain, 2010). Genetic engineering is thus a central part in the quest to establish sustainable and efficient processes for the production of fuels, chemicals, food ingredients, and pharmaceutical products.

Most of these achievements would not been possible without sequencing technologies that allowed us to identify the genetic sequences and validate the genetic manipulations in microorganisms. More recently, Next-Generation Sequencing (NGS) technologies have provided us with the capability of fast and cheap sequencing of DNA at an unprecedented scale. NGS has allowed *de novo* assembly of the genomes of thousands of organisms for which no genome sequences were previously available, ranging from complex multicellular organisms (Li et al., 2010; Nakamura et al., 2013; Pegadaraju et al., 2013; Kelley et al., 2014) to microorganisms (Soares-Castro and Santos, 2013; Yamamoto et al., 2014). NGS technologies also provide us with the means to re-sequence organisms (Atsumi et al., 2010; Wang et al., 2014), i.e., the sequencing of genetically distinct strains that are close enough to a reference strain with a sequenced genome. Re-sequencing is used to determine genetic variants ranging from single nucleotide variants (SNV) to more complex structural variants such as large deletions, inversions, and translocations. The falling cost of sequencing allows routine re-sequencing of strains isolated from the wild, monitoring the genetic stability of production strains during genetic engineering and fermentation processes, and determining the genetic basis of adaptive laboratory evolution (ALE) (Herrgård and Panagiotou, 2012). In addition to biotechnological applications, re-sequencing of microbial strains plays also a key role in other areas such as epidemiology of infectious diseases caused by bacterial and fungal pathogens, and in understanding the effects of human activity on microbial diversity and evolution in the environment.

Genome-scale metabolic models (GSMs), consisting of biochemical reactions and their relations to the genome and proteome of a cell [through gene–protein-reaction (GPR) associations], are a proven framework for the *in silico* analysis of the metabolic physiology of microbes. Genome-scale metabolic models have also been used successfully for the design of metabolically engineered strains with improved production of commercially valuable proteins and metabolites: recombinant antibodies, food additives (e.g., vanillin), organic acids, ethanol, among others (Tepper and Shlomi, 2009; Brochado et al., 2010). These models have become increasingly popular over the past decade, and more than 100 models for different organisms have been published up to this date (http://optflux.org/models). The greatest strength of GSMs lie in their simplicity and computational efficiency; new GSMs can be readily built from genomic annotations complemented with limited experimental data, and predictions from GSMs can be obtained using standard mathematical optimization methods (Varma and Palsson, 1993; Segrè et al., 2002; Shlomi et al., 2005) allowing phenotypic predictions within minutes.

Genetic variation that entails a complete loss of function – commonly referred to as gene knockout – has been successfully used to tailor GSMs to a specific genotype to improve the production of valuable compounds [e.g., biobutanol (Lee et al., 2008), sesquiterpene (Asadollahi et al., 2009), vanillin (Brochado et al., 2010), polyhydroxyalkanoates (Puchałka et al., 2008), or L-valine (Park et al., 2007)], but so far no methodological framework has been developed that would allow the incorporation of other types of genetic variants systematically. In this work, we review existing tools for analyzing genetic variants that capture more subtle changes such as synonymous and non-synonymous SNVs in coding regions or variants in promoter or other regulatory regions. We will focus on outlining the challenges of combining more subtle genetic variant information with GSMs in order to use models to predict strain-specific phenotypes.

2. UNVEILING THE EFFECTS OF GENETIC VARIATION

2.1. GENETIC VARIABILITY

Genetic variants, including SNVs and larger structural variants are commonly seen when natural or engineered strains are re-sequenced (**Figure 1**). SNVs can be found across the genome in different functional regions: (i) protein coding sequences, (ii) promoters and other regulatory elements such as ribosome binding sites, (iii) splice sites and other regions affecting transcript structures, and (iv) other genomic regions with unknown direct connections to any given protein function. Moreover, insertions or deletions of nucleotides (indels) within a coding region can cause a shift in the open reading frame usually denoted as frameshift mutations (**Figure 1A**). At the genome structure level, chromosomal rearrangements, e.g., swaps, inversions, deletions, and insertions, can affect the function of one or more proteins (**Figure 1B**).

The spectrum of the resulting effects caused by these genetic variations on individual gene or protein function or expression is very broad. Non-synonymous SNVs or in-frame indels in protein coding sequences can disrupt, enhance, or modify the activity of the protein depending on the exact amino-acid change introduced. Introduction or removal of a stop codon by specific SNVs or out-of-frame indels would be expected to result in more drastic changes of protein function. For example, the appearance of a stop codon might lead to the separation of a multi-domain protein to multiple individual single-domain proteins. The removal or replacement of a stop codon could cause translational read-through leading to an elongated protein with potential new functions (Long et al., 2003). SNVs and indels in regulatory regions such as promoters can affect the transcription or translation processes giving rise to variation in expression levels in specific proteins. In eukaryotes, variants within introns can also affect transcript structures by introducing new exons or removing existing ones. Some variations can also be completely silent with no change of phenotype, for example, a change in a stop codon location might not change the protein activity. Ideally, we should be able to predict the degree in which single and multiple genetic variants within or near a coding locus affect the relevant protein function or expression. This would allow us to rapidly make sense of the vast quantities of re-sequencing data that is becoming available without having to test the effects of all variants experimentally.

Larger-scale structural variations, such as duplications, deletions, translocations, and inversions, can have significant effects on the expression or activity of individual proteins. For example, there can be a complete loss of one or more genes, or a duplication of genomic regions can modify the expression of multiple genes within or nearby these regions (Blount et al., 2012). Very large-scale genomic changes, such as duplication of entire chromosomes, can change the activity of hundreds of proteins at once and have been reported in both natural microbial strains (Gordon et al., 2009) and in strains created by ALE (Caspeta et al., 2014). The effects of structural genomic variation are often more systemic than the effects of smaller scale variations, but any framework attempting to predict the phenotypic effects of genetic variation needs to consider both small- and large-scale variation.

2.2. *IN SILICO*: PREDICTING THE EFFECT OF GENETIC VARIANTS

A major challenge to understanding the phenotypic consequences of genetic variation lies in our ability to predict the mechanistic consequences of mutations. Proteins are very complex structures

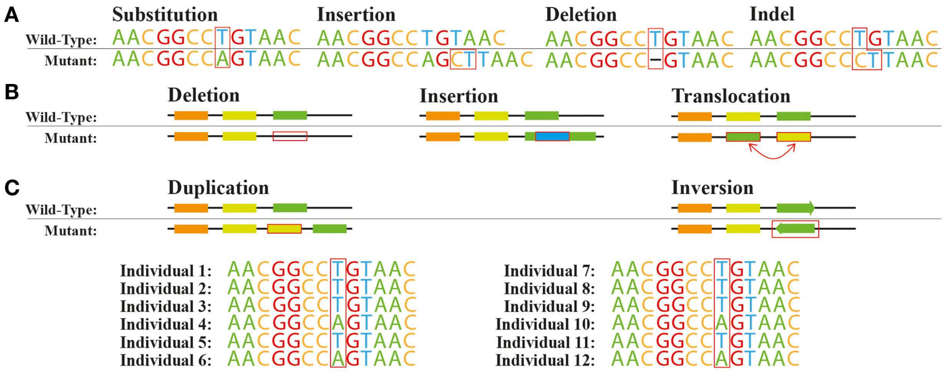

A
Substitution
Insertion
Deletion
Indel
Wild-Type:
Mutant:
AACGGCCTGTAAC
AACGGCCAGTAAC
AACGGCCTGTAAC
AACGGCCAGCTTAAC
AACGGCCTGTAAC
AACGGCC-GTAAC
AACGGCCTGTAAC
AACGGCCCTTAAC
B
Deletion
Insertion
Translocation
Wild-Type:
Mutant:
C
Duplication
Inversion
Wild-Type:
Mutant:
Individual 1: AACGGCCTGTAAC
Individual 2: AACGGCCTGTAAC
Individual 3: AACGGCCTGTAAC
Individual 4: AACGGCCAGTAAC
Individual 5: AACGGCCTGTAAC
Individual 6: AACGGCCAGTAAC
Individual 7: AACGGCCTGTAAC
Individual 8: AACGGCCTGTAAC
Individual 9: AACGGCCTGTAAC
Individual 10: AACGGCCAGTAAC
Individual 11: AACGGCCTGTAAC
Individual 12: AACGGCCAGTAAC

FIGURE 1 | Common genetic variations. Variations at the **(A)** nucleotide level and **(B)** structural level. **(C)** Single nucleotide polymorphism A/T across a population.

that fall into different functional categories and can be characterized by many distinct properties. For example, how protein activities are measured depends on their functional category: transcription factors can be characterized by their binding strength to a certain promoter region while metabolic enzymes would typically be characterized by their catalytic activity and specificity for a certain substrate. Moreover, proteins do not operate in isolation but interact with each other and with metabolites, and these interactions have consequences on the activities of proteins. Here, we provide a non-exhaustive review of the types of methods that are commonly used to predict the effects of genetic variants on protein function.

The study of single nucleotide polymorphisms (SNP) that affect human health is one of the major focus areas of modern medical research. In human genetics, SNPs are single nucleotide substitutions found in more than 1% of a population. Several algorithms were implemented to determine the effect of SNPs, mostly specialized to the analysis of human genotyping data (see **Table 1** and **Figure 2**). One limitation of most of these algorithms is that they are binary classifiers – deleterious or neutral, disease-causing or neutral, and tolerant or intolerant. This means that the genetic changes will either be predicted to have no effect or to cause some measurable, negative impact on the phenotype. This may not be an issue in the context of human diseases as SNP data are primarily used in diagnostics. However, fine tuning engineered microbial strains requires more than a black and white approach for predicting variant effects on protein function. This is because many genetic variants can yield proteins with either increased or decreased activity, requiring methods that are able to predict also potential gains or modifications of functions. In particular, when mutagenesis and selection or ALE methods are applied, one commonly sees gain of function mutations of specific genes that are crucial for the adaptation to, for example, new carbon sources (Conrad et al., 2011).

Of the existing algorithms (**Table 1**), *SIFT* (**S**orting **I**ntolerant **f**rom **T**olerant) (Ng and Henikoff, 2001) is often used as a gold standard to compare the performance of new algorithms or as a foundation for novel prediction strategies. SIFT and related approaches are based on the notion that evolutionary conservation can be used to predict the functional importance of each amino-acid in a protein and the impact of specific amino-acid substitutions. These methods typically use multiple sequence alignments of related proteins to determine a probabilistic description of what amino-acid substitutions are allowed in specific sites within the target protein. These descriptions can be used to determine the probability that non-synonymous coding SNPs observed in a re-sequencing data set will be tolerated by the protein; substitutions with a probability score smaller than a threshold are assumed to be deleterious (Kumar et al., 2009).

Sorting intolerant from tolerant provides only a binary deleterious/non-deleterious classification, and other methods have been developed to allow predicting cases where SNPs improve protein function. The *Polyphen* (Ramensky, 2002) and *PolyPhen2* (Adzhubei et al., 2010) approaches provide the means to discriminate three states when analyzing the effect of a SNP: benign, neutral, or deleterious. *Polyphen* uses a list of predetermined rules that combine the output of multiple algorithms using combinations of structural and sequence-based measures of mutation impact. *PolyPhen2* uses a machine-learning approach (a naive Bayes model) to predict an overall score for the variant effect, and the classification to three categories is based on thresholds. Although the algorithm is trained with human datasets, similar methods could potentially be used to build predictive models for variant effects in microorganisms. The overall variant effect score could also be exploited in more advanced methods that combine scores from different variants affecting different proteins to make phenotypic predictions.

Most studies on genetic variation focus on SNPs and disregard indels, which are also commonly observed when related microbial strains are compared to each other. The *PROVEAN* (Choi et al., 2012) and *Mutation taster 2* (Schwarz et al., 2014) approaches are capable of analyzing both SNPs and indels. *PROVEAN* uses substitution matrix scores (i.e., BLOSUM62) with gap and extension penalties to compute a variation score between the wild-type and mutant. More recently, *Mutation taster 2* computes several features (structural and evolutionary properties) for the mutated sequence using a Bayes classifier.

One possible approach for improving our ability to predict variant effects on protein function would be to predict effects of amino-acid changes on protein stability and folding (Khan and Vihinen, 2010). There are a number of tools available for these tasks (Khan and Vihinen, 2010), and stability predictions could be used to predict variant effects on protein function, as strongly destabilizing mutations would result in complete loss of function for the protein. Methods for predicting variant effects on protein stability have only been found to be moderately accurate in independent evaluation studies (Khan and Vihinen, 2010). For this reason, stability predictors should be combined with other variant effect prediction approaches to improve their predictive power for general variant effect analysis. The application of these types of stability prediction methods will be discussed in Section 3.2 in more detail together with the applications of metabolic modeling.

The majority of algorithms (53%) for variant effect prediction listed in **Table 1** rely on machine-learning approaches [e.g., AUTO-MUTE (Masso and Vaisman, 2010), FunSAV (Wang et al., 2012), or HANSA (Acharya and Nagarajaram, 2011)], which is a practical strategy given the huge amount of data available for human diseases. Regarding the selection of features, most methods use evolutionary conservation information (92%) and more than half rely on structural properties (69%). The selection of sufficient features is a challenge in itself; no matter what approach is used, it is necessary to define which properties and attributes of proteins are capable of discriminating the phenotypes of interest. The improvements in the prediction capabilities provided by sequence-, evolution-, or structural-based features has been previously studied, and these studies have shown that the inclusion of structural properties leads to significant improvements in predictive power (Saunders and Baker, 2002). This has been recently confirmed by a benchmark performance test that includes several of the existing algorithms (Thusberg et al., 2011). Another effort to benchmark and improve different approaches is the Critical Assessment of Genome Interpretation (CAGI) community, which organizes a benchmark competition on predicting the effect of genetic variants on known disease phenotypes.

Table 1 | A summary of the available software tools for predicting the effect of the genetic variants.

Tool	Description	Reference
AUTO-MUTE	Uses the "4-Body Statistical Potential" to compute a set of features – based on protein 3D structure – used to train a Random Forest model to predict *neutral* or *disease*-associated SNPs.	Masso and Vaisman (2010)
Align-GVGD	This algorithm is based on multiple sequence alignment and Grantham distance to identify missense SNPs. The authors propose a measure to calculate how much the substitution changes the Grantham distance.	Tavtigian (2005)
CADD	A machine-learning approach that uses a SVM model to predict deleterious phenotypes caused by SNPs.	Kircher et al. (2014)
Chasman and Adams (2001)	A probabilistic approach to identify which SNPs have an effect on the protein function using structural and evolutionary features that compare the variation against a dataset of mutations of lac repressor and T4 lysozyme.	Chasman and Adams (2001)
CONDEL	**Con**sensus **del**eteriousness provides a score computed based on the weighted average of the normalized scores of five different tools: LogR.E-value, MAPP, mutation assessor, polyphen, and STIF.	González-Pérez and López-Bigas (2011)
Evolutionary action	Evolutionary action is a function that links genotype with phenotype using evolutionary information, by quantifying the impact of SNPs on the fitness of a population; it correlates with disease-associated mutations.	Katsonis and Lichtarge (2014)
FATHMM	Uses Hidden Markov Models (HMMs) to obtain position-specific information. The prediction is based on the probability change of the HMM between wild-type and mutant.	Shihab et al. (2012)
FunSAV	A random forest classifier for predicting deleterious SNPs. It combines properties of the mutated protein with other tools (i.e., nsSNPAnalyzer, PANTHER, PhD-SNP, PolyPhen2, SIFT, and SNAP).	Wang et al. (2012)
FuzzySnps	A machine-learning approach that uses a Random Forest model trained by combining "4-Body Statistical Potential" and sequence-based features to identify tolerant and intolerant SNPs.	Barenboim et al. (2008)
Goldgar et al. (2004)	A probabilistic approach to determine if a SNP is disease-causing, which is achieved by computing the likelihood of the protein to be similar to previously classified mutated proteins in a dataset.	Goldgar et al. (2004)
HANSA	It is a machine-learning classifier that uses a SVM model to predict whether a SNP will be neutral or disease-causing.	Acharya and Nagarajaram (2011)
LogR.E-value	Uses the *E*-value computed by the HMMER algorithm using PFAM motifs to distinguish between deleterious and neutral SNPs.	Clifford et al. (2004)
LS-SNP	A workflow/database that uses predefined rules and machine-learning (SVN) approach to systematically characterize known SNPs.	Karchin et al. (2005)
Krishnan and Westhead (2003)	Two machine-learning approaches – using SVM and Decision Trees models – are used to predict the "effect" or "no-effect" of a SNP.	Krishnan and Westhead (2003)
MAPP	**M**ultivariate **A**nalysis of **P**rotein **P**olymorphism uses statistical analysis to predict the deleterious effect of SNPs.	Stone (2005)
Mutation assessor	Predicts the degree of impact in a protein by scoring the mutation based on the impact it causes regarding the properties of a multiple sequence alignment of homologous sequences.	Reva et al. (2011)
Mutation taster 2	Uses a Bayes classifier to predict disease associated effects caused by SNPs or Indels. The classifier uses a set of features that includes splicing site and polyadenylation signal information along with structural and evolutionary properties.	Schwarz et al. (2014)
MutPred	Uses a machine-learning approach to predict disease or neutral SNPs. The features used refer to a probability of loss or gain of function regarding several functional and structural properties of the encoded protein. The authors trained SVM and Random Forest models in this work.	Li et al. (2009)
nsSNPAnalyzer	Uses a Random Forest model trained with features (consisting of SIFT score and information from multiple sequence alignment and protein 3D structures) to identify disease associated SNPs.	Bao et al. (2005)
Papepro	A SVM prediction model is used by the authors to separate deleterious from neutral SNPs.	Tian et al. (2007)

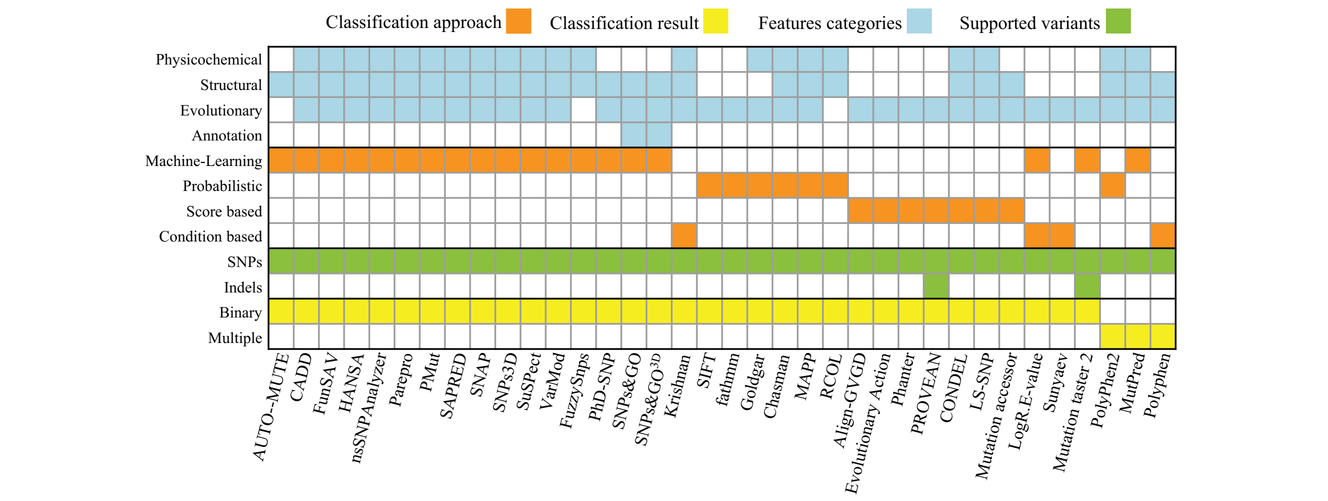
(Continued)

Table 1 | Continued

Tool	Description	Reference
Panther	Using an internal database of HMM, an evolutionary score is computed and the method predicts deleterious or neutral effects with a probability attached. The cutoff can be defined by the user (default is 3).	Thomas and Kejariwal (2004)
PhD-SNP	This approach uses one of two SVM models: one is trained using sequence profile features and the other is trained using sequence features. The choice of which model to use is based on a preliminary decision: if the mutation exists in the homology profile, the first model is used, otherwise the prediction is done using the second model.	Capriotti et al. (2006)
PMut	Predicts pathological or neutral effects of amino-acid substitutions. The prediction model is a neural network using structural-, physicochemical-, and evolutionary-based features, all calculated using sequence information only (without requiring a3D protein structure).	Ferrer-Costa et al. (2005)
Polyphen	A set of rules defined by the authors is used to predict the effect of a SNP. These rules are built based on three properties: PSIC score, substitution site properties, and substitution type properties. If one of the rules matches, the output can be deleterious or benign, otherwise the substitution is classified as neutral.	Ramensky (2002)
PolyPhen2	The follow up version of Polyphen, uses a naive Bayes predictor to predict damaging, benign, or neutral effects of SNPs. It uses structural information if available.	Adzhubei et al. (2010)
PROVEAN	**Pro**tein **V**ariation **E**ffect **AN**alyzer computes a score based on evolutionary information to predict if a genetic variant (i.e., SNP or Indel) is neutral or deleterious.	Choi et al. (2012)
RCOL	Applies a Bayes' formula to calculate the probability of a SNP to be deleterious. The likelihood is tested using 20 structural and physicochemical parameters.	Terp et al. (2002)
SAPRED	Using a SVM prediction model, the authors combine features computed from evolutionary, structural, and physicochemical properties to predict disease associated SNPs.	Ye et al. (2007)
SIFT	Using a PSSM, SIFT determines the probability of a substitution being tolerated in a given position.	Ng and Henikoff (2001)
SNAP	Identifies non-neutral SNPs using machine-learning approaches that combines a battery of Neural Network models.	Bromberg et al. (2008)
SNPs3D	Combines a set of features obtained from protein 3D structure and evolutionary information to predict deleterious effects using a SVM model.	Yue et al. (2006)
SNPs&GO	A machine-learning approach that includes GO annotations as features in a SVM model to predict whether a SNP is neutral or disease associated.	Calabrese et al. (2009)
SNPs&GO^{3D}	It is the successor of SNPs&GO. It includes new features obtained from protein 3D structure.	Capriotti and Altman (2011)
Sunyaev (2001)	This approach uses a set of seven rules empirically defined by the authors to identify nsSNPs. If one of the rules is matched, then the SNP is likely to be deleterious.	Sunyaev (2001)
SuSPect	A SVM model implementation to predict disease phenotypes caused by SNPs. The authors started with a high number of features until they identified nine that provided best performance.	Yates et al. (2014)
VarMode	A machine-learning approach using a SVN model to predict the effect of SNPs that includes information regarding known protein–protein interactions. It predicts non-synonymous SNPs.	Pappalardo and Wass (2014)

While the majority of algorithms aim to predict variant effects on individual proteins, a different objective is followed by the SNP-IN method that predicts how protein–protein interactions (PPIs) are affected by a SNP (Zhao et al., 2014). This is achieved by a set of features that includes the relative free energy change between wild-type and mutant PPI, the energy of all interactions in a protein complex, and other physicochemical properties, e.g., hydrophobic solvation or water bridges. Using these features, supervised and semi-supervised machine-learning approaches are used to predict how deleterious SNPs are. This approach is a very interesting, as changes in PPIs could be used to explain epistatic interactions between multiple variants. Like some previously mentioned prediction algorithms, SNP-PI requires an existing 3D model of the protein structure and, in addition, knowledge of the PPIs a given protein is involved in.

At a larger scale, genome-wide association studies are used to identify how differences between hundreds of thousands of individuals and make genotype to phenotype consequences. This approaches work as black boxes and make use of statistical and machine-learning approaches that require huge datasets. The



Classification approach 🟧 Classification result 🟨 Features categories 🟦 Supported variants 🟩

	AUTO--MUTE	CADD	FunSAV	HANSA	nsSNPAnalyzer	Parepro	PMut	SAPRED	SNAP	SNPs3D	SuSPect	VarMod	FuzzySnps	PhD-SNP	SNPs&GO	SNPs&GO^{3D}	Krishnan	SIFT
Physicochemical		🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦				🟦	
Structural	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	
Evolutionary		🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦		🟦	🟦	🟦	🟦	🟦
Annotation															🟦	🟦		
Machine-Learning	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧		
Probabilistic																		🟧
Score based																		
Condition based																	🟧	
SNPs	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩
Indels																		
Binary	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨
Multiple																		

	fathmm	Goldgar	Chasman	MAPP	RCOL	Align-GVGD	Evolutionary Action	Phanter	PROVEAN	CONDEL	LS-SNP	Mutation accessor	LogR.E-value	Sunyaev	Mutation taster 2	PolyPhen2	MutPred	Polyphen
Physicochemical		🟦	🟦	🟦	🟦					🟦	🟦					🟦	🟦	
Structural			🟦	🟦	🟦					🟦	🟦	🟦				🟦	🟦	🟦
Evolutionary	🟦	🟦	🟦	🟦		🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦
Annotation																		
Machine-Learning													🟧		🟧		🟧	
Probabilistic	🟧	🟧	🟧	🟧	🟧											🟧		
Score based						🟧	🟧	🟧	🟧	🟧	🟧	🟧						
Condition based													🟧	🟧				🟧
SNPs	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩
Indels									🟩						🟩			
Binary	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨			
Multiple																🟨	🟨	🟨

FIGURE 2 | Summary of properties and approaches for software listed in Table 1. The approaches found fall into four different categories: *Machine-Learning*, *Probabilistic*, *Score* (calculating a summarizing score of a set of hand-picked statistics), and *Rule* (using a set of empirically derived rules). These approaches provide one of two types of classifications each: a binary classification (e.g., neutral or deleterious) or a multi-classification (e.g., benign, neutral, and deleterious). The features used by those approaches can be computed based on properties of the following five categories: (i) physicochemical properties (e.g., solvent accessibility, polarity, charge, disorder, and Grantham), (ii) structural information about the primary, secondary, and tertiary structure of a protein (e.g., α-helices, β-sheets, and coil), (iii) evolutionary properties (multiple sequence alignments, position-specific scoring matrices, and Hidden Markov models), and (iv) genome annotation (GO terms or other protein function annotations). The supported variants were determined either by accessing the tools' websites or by the description of the approach itself.

current work and applications (e.g., clinical risk assessment) have been recently reviewed (Okser et al., 2014).

2.3. *IN VIVO*: DEEP MUTATIONAL SCANNING AND TN-SEQ

Next-generation sequencing has enabled studying the effects of genetic variation on individual proteins or regulatory elements *in vivo* and *in vitro*. Deep mutational scanning (DMS) is an effective high-throughput method to measure the effects of mutations on protein stability and function (Fowler and Fields, 2014). The space of all possible amino-acid substitutions in a protein is exhaustively screened by first constructing a library of sequence variants using standard techniques like error prone PCR, then by using a high-throughput assay to select variants based on a fitness measure (e.g., growth rate, ligand binding, or product fluorescence), and finally by applying deep sequencing to the selected and unselected sequence variant pools. This approach results in a matrix that contains fitness values for each amino-acid substitution discovered in the selected pool. Depending on the method used for creating sequence diversity and sequencing depth, DMS can also be used to measure epistatic effects between substitutions at different sites.

The applicability of DMS is primarily limited by the lack of high-throughput functional assays for most proteins and, so far, DMS has not been applied to metabolic enzymes. When DMS can be applied at a broader scale, the results obtained from the assay could increase the predictive power of bioinformatic tools for genetic variation analysis by providing more complete training datasets for the types of predictive methods discussed in the previous section. Methods similar to DMS can also be used to systematically study effects of genetic variation in regulatory regions on protein expression using fluorescence protein-based assays.

Here, we will highlight a few case studies using DMS and related methods to study protein or regulatory element function. In the analysis of *Saccharomyces cerevisiae* poly(A)-binding protein (Melamed et al., 2013), strong epistatic effects between substitutions at specific sites were discovered. Although epistasis was not widespread, this is worrying from a computational modeling perspective, as modeling approaches usually do not account for epistasis. Another important highlight is the identification of alternative start codons. Although analyzed in previous studies, the DMS has shown that some amino-acids can be replaced by methionine and yield functional proteins (Kim et al., 2013). This biological information can be extrapolated to other studies and is highly relevant when developing strategies to understand the effect of mutations, either *in vivo* or *in silico*. Strategies similar to DMS have also been used to systematically study the effects of variation in transcription factor binding sites and other regulatory elements such as ribosomal binding sites (Kosuri et al., 2013). These studies will build the foundation for predicting effects of non-coding sequence variants on protein expression.

The methods described above allow us to systematically study the effects of a large number of variants in individual proteins or regulatory regions. In microorganisms, it is also possible to use a next-generation sequencing-based method called Tn-seq to systematically study the effect of disruption of a large number of genomic loci on cellular phenotypes (van Opijnen and Camilli, 2013). Transposons are mobile DNA elements that can disrupt a genetic locus by integrating themselves into it (**Figure 1B**). Tn-seq, using high density transposon insertion libraries, can be used to interrogate the function of, for example, regulatory elements and specific protein domains in a single genome-wide assay (van Opijnen and Camilli, 2013). Tn-seq has found many applications in microbiology, and it has been used for the identification of gene function, understanding genome organization, mapping genetic interactions, or assessing gene essentiality (van Opijnen and Camilli, 2013; Yang et al., 2014). Tn-seq does not offer a resolution on the single base-pair level, but the method can be rapidly used to generate sub-gene-level information relating, for example, to the essentiality of specific domains in a protein. This information in turn could be used to improve variant effect predictions, as variants in essential domains of a protein would be more likely to be predicted to be deleterious than variants in non-essential domains of the same protein.

3. PREDICTING PHENOTYPES FROM GENOTYPES AT THE GENOME-SCALE

3.1. STATISTICAL AND NETWORK-ORIENTED APPROACHES FOR PREDICTING PHENOTYPES FROM GENOTYPES

Section 2 focused on the task of predicting the effects of genetic variation on individual protein function or expression. However, this is only a small part of a much larger problem, which of predicting cellular or organism phenotypic effects of all the genetic variants present in a genome. This requires combing the effects of variation on the function and expression of all proteins. So far, there have been surprisingly few efforts to take all genetic variants discovered in an individual (either a human or a microbial strain) and attempt to predict how certain phenotypes would be affected by all these variants together (Burga and Lehner, 2013; Lehner, 2013).

One of the first systematic attempts toward this goal was the pioneering study by Jelier et al. in *S. cerevisiae*, where growth phenotypes of selected yeast strains under different conditions were predicted from genetic differences between a reference strain and the strain of interest (Jelier et al., 2011). This was achieved by first predicting effects of coding and regulatory variants on protein function and expression using approaches similar to the one outlined in the previous section. These variant effect predictions were then combined into a single phenotypic prediction for the strain, using published single gene deletion growth phenotyping data for a yeast reference strain under the same condition. This approach can be considered to be highly simplistic, as the effects of multiple genetic variants acting on separate proteins were treated cumulative. Despite this, the approach still allowed accurate prediction of growth phenotypes across a broad range of conditions. There have also been a number of other approaches for predicting broader phenotypic consequences of single variants by mapping the variant data onto biological networks such as PPI or genetic networks (Carter et al., 2013). However, these approaches have typically not attempted to use the whole genotype of an individual (i.e., more than one variant at a time) to predict specific phenotypes.

3.2. USING GENOME-SCALE METABOLIC MODELS FOR INTERPRETING GENETIC VARIANTS

The phenotype prediction methods described above are data-driven and use statistical models to predict the effects of genetic variants in the context of biological networks. However, for metabolic networks we can go beyond statistical models and graph-based descriptions to constraint-based models that are scalable to the genome-level and incorporate physicochemical, flux capacity, and reaction directionality constraints [see Price et al. (2004) for a review of constraint-based modeling]. This type of mechanistic modeling approach is very useful for understanding genetic changes that affect specific metabolic phenotypes. For example, the study of SNPs that affect mitochondrial metabolism (Jamshidi and Palsson, 2006) is a good example of how variant data can be mapped onto metabolic networks in order to explain the mechanistic basis of disease phenotypes.

A genome-scale metabolic models are composed of biochemical reactions, collected from literature and the genome annotation of an organism. This system of reactions is encoded as a matrix of stoichiometric coefficients that is usually referred to as stoichiometry matrix[1]. Assuming metabolism is in a steady-state, i.e., metabolite concentrations do not change over time, all fluxes have to balance each other. These flux-balances constitute linear constraints that can easily be analyzed using methods from linear algebra.

Furthermore, after inclusion of further constraints, e.g., known uptake and secretion rates and knowledge about reaction directionality, linear optimization methods can compute biologically relevant flux vectors that maximize defined objective functions. For example, growth can be simulated by maximizing the consumption of biomass precursors in empirically determined proportions. This type of analysis is usually referred to as flux balance analysis [FBA; see Orth et al. (2010) for a comprehensive introduction to this method].

Global optimal solutions to this linear optimization problems can be calculated very efficiently using linear programing (computation times are on a millisecond to second range for genome-scale models). Thus, one can compute thousands of phenotypes in a few minutes, simply by changing the constraints of the problem [see Lewis et al. (2012) for a comprehensive list of available *in silico* methods and (Bordbar et al., 2014) for a review of their applications].

Since the relationship between reactions, enzymes, and genes (usually referred to as GPR associations) is usually known and encoded in these models, the effect of a gene knockout can readily be mapped to the associated reactions by constraining their fluxes to be zero or by removal from the model. This way FBA can be used to compute the metabolic phenotype associated with a metabolic gene deletion, making it suitable for the analysis of genetic variation data that involves deletions or other mutations that lead to the complete loss of function of enzymes.

Flux balance analysis assumes that knockout strains can recover to an optimal growth phenotype, which might be unrealistic in cases where regulatory mechanisms – not modeled explicitly in these models – might not be able to accommodate the desired state. Other methodologies [e.g., ROOM (Shlomi et al., 2005), MoMA (Segrè et al., 2002), MiMBl (Brochado et al., 2012), and RELATCH (Kim and Reed, 2012)] employ more plausible assumptions and have been shown to improve the accuracy of knockout

[1] The rows and columns of the stoichiometry matrix correspond to metabolites and reactions respectively; negative (positive) factors represent consumption (production) of substrates (products).

predictions. For example, MoMA minimizes the euclidean distance of the wild-type and mutant flux distributions, assuming that a mutant reaches the closest feasible flux distribution that is not necessarily optimal. The predictive power of FBA and these other approaches have been extensively assessed using genome-wide gene knockout assays (Snitkin et al., 2008) and transposon insertion libraries (Yang et al., 2014) and have resulted generally in a high degree of accuracy (Monk and Palsson, 2014).

Constraint-based models have also been applied to predict epistatic interactions by simulating effects of pairwise gene deletions, but with a significantly reduced accuracy in comparison to single deletions (Szappanos et al., 2011). Furthermore, simulations of multiple gene deletions have been successfully applied in developing design strategies for metabolic engineering by redirecting flux to desired products (Milne et al., 2009; Blazeck and Alper, 2010).

A number of limiting factors can diminish the ability of constraint-based models to predict phenotypic effects of loss of function mutations: (i) missing reactions and erroneous GPRs, (ii) erroneous flux constraints due to the lack of thermodynamic or regulatory information, and (iii) the assumption of a fixed biomass composition that is known to change across growth conditions. Even with these limitations, constraint-based models still outperform statistical models in predicting consequences of gene deletions (Szappanos et al., 2011).

Since constraint-based models have demonstrated good ability to predict phenotypic outcomes of single and multiple gene deletions, these models should also be useful for predicting effects of other genetic variants. A SNV or indel that is predicted to reduce the maximal flux rate of an enzyme can be used to constrain the upper bound of a flux. FBA and similar methods can be used to compute the effects of these variations on the phenotype, providing a system-wide overview of the effects caused by the substitution (Jamshidi et al., 2007). This is a fast and effective way of predicting phenotypes, but it requires that one can estimate the effect the variant has on the maximum flux rate. Nevertheless, cases of complete loss of function fall into the same category as gene knockouts, and combining the bioinformatic prediction tools discussed in Section 2.2 with modeling capabilities can be used to integrate variant data. This approach can also be extended to any number of variants and genes, with the caveat that epistatic interactions are currently not captured accurately by the models.

There is currently only a limited number of studies that use GSMs to systematically explore the effects of genetic variants on phenotypes. Chang et al. (2013) conducted a study where GSMs coupled with protein structures of metabolic enzymes (GEM-PRO[2]) were used to interpret genetic variant data of *Escherichia coli* strains evolved to tolerate high temperatures (Chang et al., 2013). In this study, a GSM of *E. coli* was constrained using experimentally or bioinformatically determined thermostabilities of metabolic enzymes. Since the maximum flux capacity of a reaction is proportional to the concentration of active enzyme, temperature changes can be modeled by varying the flux constraints accordingly. This enables the prediction of enzymatic steps that are disproportionately temperature sensitive. For the evolved strains, flux balance analysis was used to explore the adaptation of the mutated enzymes; constraints associated with mutated proteins were relaxed to explain the experimentally measured growth rates (Chang et al., 2013). The study did not include separate predictions of variant effects on protein function, but rather treated all variants observed in a protein as potentially affecting its activity.

A more recent study by Nam et al. (2014) describes the use of GSMs for understanding the metabolic effects of cancer mutations. In particular, Nam et al. use genetic mutation information, gene expression profile data, and a human GSM (Thiele et al., 2013) to construct context-specific models for different cancer types. Loss and gain of function were systematically analyzed. Loss of function was modeled as described above (i.e., constraining affected reactions' fluxes to 0). Gain of a function, on the other hand, was modeled by adding novel promiscuous activities as predicted by chemoinformatic approaches. This approach allowed the prediction of potential oncometabolites.

3.3. KINETIC MODELING OF GENETIC VARIANTS

As mentioned in the previous section, constraint-based modeling does not provide any information about the dynamic behavior of a metabolic system. A full kinetic description of a biochemical reaction network can be formulated using ordinary differential equations (Heinrich and Schuster, 1996). The major advantage of using kinetic models to study effects of genetic variation lies in their ability to account for mutations affecting catalytic or regulatory sites of an enzyme, causing either a gain or loss of catalytic activity, or binding sites of allosteric regulators.

Previous studies of red blood cell metabolism provide an overview on how SNPs can alter kinetic parameters and how kinetic models can be used to explain metabolic syndromes caused by enzyme deficiencies (Jamshidi, 2002; Jamshidi and Palsson, 2009). A disadvantage of using kinetic models is that kinetic parameters are not available for most enzymes and measuring the parameters can be challenging. For this reason, building predictive genome-scale kinetic models remains a challenge (Stanford et al., 2013). Kinetic models are a viable tool for interpreting genetic variant data only in specific cases like, for example, the red blood cell that harbors a relatively simple metabolism.

4. CONSIDERATIONS AND FUTURE DIRECTIONS

4.1. METHODS AND TOOLS TO PREDICT THE EFFECT OF GENETIC VARIANTS

Many approaches have been explored in the past decade to understand and analyze the effects of genetic variation. In particular, the most active field has been the application of NGS techniques to characterize of genetic variation in the context of human disease. The amount of disease related information makes machine-learning approaches very suitable for the purpose of predicting effects of single genetic variants. Since most prediction methods have been trained and tested with human data, many of the existing methods do not perform as well or are simply not suited for the analysis of microbial genetic variants.

The other area where the study of microbial genetic variation lags behind human genetics is the systematic collection of variant and phenotyping data. Efforts to collect human genotype and

[2] Genome-scale metabolic models are sometimes also referred to as GEMs.

phenotype data in a standardized way are currently underway with databases such as dbSNP and European Variation Archive. The UniProt database also collects variants found in the proteins sequences when this information is available. Every day thousands of new environmental or pathogenic isolates and laboratory developed microbial strains are sequenced around the world, but there is no centralized repository for this data in common use. We argue that it is of utmost importance to collect genetic variant data together with associated phenotypic data in a standard way for microbes as well.

All the existing algorithms for variant effect prediction are used to classify variants to preassigned categories (for example deleterious or non-deleterious). The approaches that predict deleterious effects can already be handled as knockouts in modeling their phenotypic effects using GSMs, but more subtle effects of mutations are missed by this approach. In order to improve our ability to predict phenotypes, there is a need to move beyond classification toward quantitative measures of variant effects on individual protein function. There are numerous features related to protein function that may be relevant for predicting variant effects: evolutionary and conservation, physicochemical (e.g., charge, polarity, or free energy), and structural (e.g., secondary structures, spatial distances between amino-acids or B-factors).

Existing methods for predicting variant effects have been primarily focused on generic predictors for all proteins irrespective of their function (e.g., enzymes, transcription factors, transporters, chaperons, etc.) and how do they behave in their environment (i.e., interaction with other elements: proteins, metabolites, DNA, etc.). This limits the predictive power of the methods in cases where additional information is readily available such as the relatively well studied field of microbial metabolism. For example, for metabolic enzymes, information on how kinetic parameters are affected by mutations and how these parameters vary between enzymes from different species is systematically collected in databases such as BRENDA. This type of information could be used to build improved variant effect predictors specifically for metabolic enzymes.

4.2. MODELING AND HIGH-THROUGHPUT DATA ANALYSIS

Improvements in genome-wide variant effect prediction can also come from improving or extending genome-scale modeling approaches. Recent innovations like GEM-PRO, as discussed in Section 3.2, fulfill the requirement of 3D protein structures to predict the effects of genetic variation at the protein level and could be used to systematically analyze the effect of genetic variation on a genome-scale for metabolism.

Approximately 10–30% of the genes encoded in a microbial genome are represented in metabolic GSMs, limiting the utility of these models for interpreting genomic variant data. Metabolic GSMs can be extended in a number of ways to increase coverage of the overall set of genes. The transcriptional regulatory network represented as interactions between transcription factors and target genes, can help extend the coverage of predictive models and can be integrated with metabolic GSMs in a number of ways (Covert et al., 2004; Chandrasekaran and Price, 2010). These integrated models have been successfully used to make phenotypic predictions.

Another recent extension of GSMs is ME-Models[3]. These models account for the entire machinery needed for gene and protein expression, providing a higher coverage of cellular functions and a higher resolution of cellular composition (O'Brien et al., 2013). ME-models have also been extended further to incorporate protein translocation from the cytoplasm to the periplasm (Liu et al., 2014). Currently, most of these extensions of GSMs have only been developed for *E. coli* and significant efforts will be required to build these extended models for other bacteria as well as eukaryotic model organisms such as *S. cerevisiae.*

The development of accurate kinetic models of metabolism, which could be useful for investigating the effects of mutations on allosteric regulation and catalytic activity, is still a tedious process. These models are usually limited to small parts of metabolism focusing on central carbon metabolism (Chassagnole et al., 2002; Peskov et al., 2012; Machado et al., 2014). There are two main reasons for these limitations: the models become huge in size and kinetic information of many enzymes is still unknown. Protocols (Stanford et al., 2013) and methodologies (Chowdhury et al., 2014) are being developed to bring kinetic modeling to the genome-scale, but the resulting models have not yet reached sufficiently mature stage for use in variant effect prediction.

In comprehensive level, a strategy for building whole-cell models by combining multiple individual models of different cellular processes including cell cycle, metabolism, transcription, and transport has been proposed (Karr et al., 2012). This strategy that also allows combining models using different representations (constraint-based, kinetic, and stochastic) was used to build a functioning whole-cell model of one of the simplest prokaryotes, *Mycoplasma genitalium.* Efforts toward building more complete genome-scale models of microbes will continue as more and more information is collected and computing power increases. These models will bring us closer to the goal of genome-wide prediction of phenotypes from genotyping data.

4.3. OPPORTUNITIES

Genetic engineering tools, such as MAGE (Wang et al., 2009) or CRISPR/Cas9 (Xu et al., 2014), already allow us to quickly edit genomes in a precise and accurate fashion at the single base-pair resolution level at multiple loci simultaneously. These methods will allow us to map epistatic interactions of variants within a single gene and between multiple genes more comprehensively than before. On the other hand, new *in silico* tools for predicting variant effects on phenotypes outlined above open the way to a new style of modeling at the scale of single nucleotides. These new modeling tools will greatly benefit from better training datasets that can be obtained using MAGE, CRISPR/Cas9 or other genome editing methods systematically to map epistatic interactions. The application of these novel strategies provides a way to fine tune activities of proteins in the context of complete cellular networks. For example, we envision that in the future we will have predictive models of how engineering of multiple enzymes at the single amino-acid level would affect the production of a desired metabolite.

To achieve the maximum potential of genome-scale biochemical network modeling and genetic variant analysis, a link must

[3] Metabolism and Expression models.

be created between these two fields. The necessary information to connect both worlds is already there: we know the genes, the proteins, and the reactions. The major limitations are in the current methods and data sources. On the one hand, we must overcome the limitations of the tools available to predict variant effects by allowing more fine grained predictions of how a variant may affect any given protein function or expression. The usage of protein folding predictions, for example, has already been established in metabolic modeling (Chang et al., 2013), and it should be possible to use tools that predict variant effects on protein stability together with genome-scale models. On the other hand, we need to improve biochemical network modeling techniques: this is a evolving field and in the past decade there have been efforts to standardize the construction of models (Thiele and Palsson, 2010) and improving prediction methods by including high-throughput data (Machado and Herrgård, 2014).

Finally, it should be acknowledged that there will always be limitations in using solely genomic variant data as the basis for making phenotypic predictions for specific strains. We may also need to measure intermediate phenotypes such as transcript, protein, or metabolite levels for these strains in order to make predictions of how a given genotype affects a specific phenotype (Burga and Lehner, 2013). Fortunately enough comprehensive multi-omic datasets are currently being collected for wild-type microbial strains, allowing refinement of modeling and bioinformatic approaches for phenotypic prediction (Ishii et al., 2007; Skelly et al., 2013). Hopefully, systematizing such datasets and a concerted action between modelers, geneticists, microbiologists, and bioinformaticians will allow us to achieve the prediction of changed and novel metabolic capabilities of a microbial strain from genomic re-sequencing data.

ACKNOWLEDGMENTS

JC, MH, and NS acknowledge support by the Novo Nordisk Foundation through the Novo Nordisk Foundation Center for Biosustainability. MA acknowledges funding from a Biotechnology-based Synthesis and Production Research grant from the Novo Nordisk Foundation.

REFERENCES

Acharya, V., and Nagarajaram, H. A. (2011). Hansa: an automated method for discriminating disease and neutral human nsSNPs. *Hum. Mutat.* 33, 332–337. doi:10.1002/humu.21642

Adrio, J.-L., and Demain, A. L. (2010). Recombinant organisms for production of industrial products. *Bioeng. Bugs* 1, 116–131. doi:10.4161/bbug.1.2.10484

Adzhubei, I. A., Schmidt, S., Peshkin, L., Ramensky, V. E., Gerasimova, A., Bork, P., et al. (2010). A method and server for predicting damaging missense mutations. *Nat. Methods* 7, 248–249. doi:10.1038/nmeth0410-248

Asadollahi, M. A., Maury, J., Patil, K. R., Schalk, M., Clark, A., and Nielsen, J. (2009). Enhancing sesquiterpene production in *Saccharomyces cerevisiae* through in silico driven metabolic engineering. *Metab. Eng.* 11, 328–334. doi:10.1016/j.ymben.2009.07.001

Atsumi, S., Wu, T.-Y., Machado, I. M. P., Huang, W.-C., Chen, P.-Y., Pellegrini, M., et al. (2010). Evolution genomic analysis, and reconstruction of isobutanol tolerance in *Escherichia coli*. *Mol. Syst. Biol.* 6, 449. doi:10.1038/msb.2010.98

Bao, L., Zhou, M., and Cui, Y. (2005). nsSNPAnalyzer: identifying disease-associated nonsynonymous single nucleotide polymorphisms. *Nucleic Acids Res.* 33, W480–W482. doi:10.1093/nar/gki372

Barenboim, M., Masso, M., Vaisman, I. I., and Jamison, D. C. (2008). Statistical geometry based prediction of nonsynonymous SNP functional effects using random forest and neuro-fuzzy classifiers. *Proteins* 71, 1930–1939. doi:10.1002/prot.21838

Blazeck, J., and Alper, H. (2010). Systems metabolic engineering: genome-scale models and beyond. *Biotechnol. J.* 5, 647–659. doi:10.1002/biot.200900247

Blount, Z. D., Barrick, J. E., Davidson, C. J., and Lenski, R. E. (2012). Genomic analysis of a key innovation in an experimental *Escherichia coli* population. *Nature* 489, 513–518. doi:10.1038/nature11514

Bordbar, A., Monk, J. M., King, Z. A., and Palsson, B. O. (2014). Constraint-based models predict metabolic and associated cellular functions. *Nat. Rev. Genet.* 15, 107–120. doi:10.1038/nrg3643

Brochado, A. R., Andrejev, S., Maranas, C. D., and Patil, K. R. (2012). Impact of stoichiometry representation on simulation of genotype-phenotype relationships in metabolic networks. *PLoS Comput. Biol.* 8:e1002758. doi:10.1371/journal.pcbi.1002758

Brochado, A. R., Matos, C., Møller, B. L., Hansen, J., Mortensen, U. H., and Patil, K. R. (2010). Improved vanillin production in baker's yeast through in silico design. *Microb. Cell Fact.* 9, 84. doi:10.1186/1475-2859-9-84

Bromberg, Y., Yachdav, G., and Rost, B. (2008). SNAP predicts effect of mutations on protein function. *Bioinformatics* 24, 2397–2398. doi:10.1093/bioinformatics/btn435

Burga, A., and Lehner, B. (2013). Predicting phenotypic variation from genotypes phenotypes and a combination of the two. *Curr. Opin. Biotechnol.* 24, 803–809. doi:10.1016/j.copbio.2013.03.004

Calabrese, R., Capriotti, E., Fariselli, P., Martelli, P. L., and Casadio, R. (2009). Functional annotations improve the predictive score of human disease-related mutations in proteins. *Hum. Mutat.* 30, 1237–1244. doi:10.1002/humu.21047

Capriotti, E., and Altman, R. B. (2011). Improving the prediction of disease-related variants using protein three-dimensional structure. *BMC Bioinformatics* 12(Suppl. 4):S3. doi:10.1186/1471-2105-12-S4-S3

Capriotti, E., Calabrese, R., and Casadio, R. (2006). Predicting the insurgence of human genetic diseases associated to single point protein mutations with support vector machines and evolutionary information. *Bioinformatics* 22, 2729–2734. doi:10.1093/bioinformatics/btl423

Carter, H., Hofree, M., and Ideker, T. (2013). Genotype to phenotype via network analysis. *Curr. Opin. Genet. Dev.* 23, 611–621. doi:10.1016/j.gde.2013.10.003

Caspeta, L., Chen, Y. P., Ghiaci, A. F., Buskov, S., Hallstrom, B. M., Petranovic, D., et al. (2014). Altered sterol composition renders yeast thermotolerant. *Science* 346, 75–78. doi:10.1126/science.1258137

Chandrasekaran, S., and Price, N. D. (2010). Probabilistic integrative modeling of genome-scale metabolic and regulatory networks in *Escherichia coli* and *Mycobacterium tuberculosis*. *Proc. Natl. Acad. Sci. U.S.A.* 107, 17845–17850. doi:10.1073/pnas.1005139107

Chang, R. L., Andrews, K., Kim, D., Li, Z., Godzik, A., and Palsson, B. Ø (2013). Structural systems biology evaluation of metabolic thermotolerance in *Escherichia coli*. *Science* 340, 1220–1223. doi:10.1126/science.1234012

Chasman, D., and Adams, R. (2001). Predicting the functional consequences of non-synonymous single nucleotide polymorphisms: structure-based assessment of amino acid variation. *J. Mol. Biol.* 307, 683–706. doi:10.1006/jmbi.2001.4510

Chassagnole, C., Noisommit-Rizzi, N., Schmid, J. W., Mauch, K., and Reuss, M. (2002). Dynamic modeling of the central carbon metabolism of *Escherichia coli*. *Biotechnol. Bioeng.* 79, 53–73. doi:10.1002/bit.10288

Choi, Y., Sims, G. E., Murphy, S., Miller, J. R., and Chan, A. P. (2012). Predicting the functional effect of amino acid substitutions and indels. *PLoS ONE* 7:e46688. doi:10.1371/journal.pone.0046688

Chowdhury, A., Zomorrodi, A. R., and Maranas, C. D. (2014). k-OptForce: integrating kinetics with flux balance analysis for strain design. *PLoS Comput. Biol.* 10:e1003487. doi:10.1371/journal.pcbi.1003487

Clifford, R. J., Edmonson, M. N., Nguyen, C., and Buetow, K. H. (2004). Large-scale analysis of non-synonymous coding region single nucleotide polymorphisms. *Bioinformatics* 20, 1006–1014. doi:10.1093/bioinformatics/bth029

Conrad, T. M., Lewis, N. E., and Palsson, B. Ø (2011). Microbial laboratory evolution in the era of genome-scale science. *Mol. Syst. Biol.* 7, 509–509. doi:10.1038/msb.2011.42

Covert, M. W., Knight, E. M., Reed, J. L., Herrgård, M., and Palsson, B. Ø (2004). Integrating high-throughput and computational data elucidates bacterial networks. *Nature* 429, 92–96. doi:10.1038/nature02456

Ferrer-Costa, C., Gelpi, J. L., Zamakola, L., Parraga, I., de la Cruz, X., and Orozco, M. (2005). PMUT: a web-based tool for the annotation of pathological mutations on proteins. *Bioinformatics* 21, 3176–3178. doi:10.1093/bioinformatics/bti486

Fowler, D. M., and Fields, S. (2014). Deep mutational scanning: a new style of protein science. *Nat. Methods* 11, 801–807. doi:10.1038/nmeth.3027

Goldgar, D., Easton, D., Deffenbaugh, A., Monteiro, A., Tavtigian, S., and Couch, F. (2004). Integrated evaluation of DNA sequence variants of unknown clinical significance: application to BRCA1 and BRCA2. *Am. J. Hum. Genet.* 75, 535–544. doi:10.1086/424388

González-Pérez, A., and López-Bigas, N. (2011). Improving the assessment of the outcome of nonsynonymous SNVs with a consensus deleteriousness score, condel. *Am. J. Hum. Genet.* 88, 440–449. doi:10.1016/j.ajhg.2011.03.004

Gordon, J. L., Byrne, K. P., and Wolfe, K. H. (2009). Additions, losses, and rearrangements on the evolutionary route from a reconstructed ancestor to the modern *Saccharomyces cerevisiae* genome. *PLoS Genet.* 5:e1000485. doi:10.1371/journal.pgen.1000485

Heinrich, R., and Schuster, S. (1996). *The Regulation of Cellular Systems*. Boston, MA: Springer. doi:10.1007/978-1-4613-1161-4

Herrgård, M., and Panagiotou, G. (2012). Analyzing the genomic variation of microbe cell fact in the era of new biotechnology. *Comput. Struct. Biotechnol. J.* 3, 1–8. doi:10.5936/csbj.201210012

Ishii, N., Nakahigashi, K., Baba, T., Robert, M., Soga, T., Kanai, A., et al. (2007). Multiple high-throughput analyses monitor the response of *E. coli* to perturbations. *Science* 316, 593–597. doi:10.1126/science.1132067

Jamshidi, N. (2002). In silico model-driven assessment of the effects of single nucleotide polymorphisms (SNPs) on human red blood cell metabolism. *Genome Res.* 12, 1687–1692. doi:10.1101/gr.329302

Jamshidi, N., and Palsson, B. Ø. (2006). Systems biology of SNPs. *Mol. Syst. Biol.* 2, 38. doi:10.1038/msb4100077

Jamshidi, N., and Palsson, B. Ø. (2009). Using in silico models to simulate dual perturbation experiments: procedure development and interpretation of outcomes. *BMC Syst. Biol.* 3:44. doi:10.1186/1752-0509-3-44

Jamshidi, N., Vo, T. D., and Palsson, B. Ø. (2007). In silico analysis of SNPs and other high-throughput data. *Methods Mol. Biol.* 366, 267–285. doi:10.1007/978-1-59745-030-0_15

Jelier, R., Semple, J. I., Garcia-Verdugo, R., and Lehner, B. (2011). Predicting phenotypic variation in yeast from individual genome sequences. *Nat. Genet.* 43, 1270–1274. doi:10.1038/ng.1007

Karchin, R., Diekhans, M., Kelly, L., Thomas, D. J., Pieper, U., Eswar, N., et al. (2005). LS-SNP: large-scale annotation of coding non-synonymous SNPs based on multiple information sources. *Bioinformatics* 21, 2814–2820. doi:10.1093/bioinformatics/bti442

Karr, J. R., Sanghvi, J. C., Macklin, D. N., Gutschow, M. V., Jacobs, J. M., Bolival, B., et al. (2012). A whole-cell computational model predicts phenotype from genotype. *Cell* 150, 389–401. doi:10.1016/j.cell.2012.05.044

Katsonis, P., and Lichtarge, O. (2014). A formal perturbation equation between genotype and phenotype determines the Evolutionary Action of protein-coding variations on fitness. *Genome Res.* 24, 2050–2058. doi:10.1101/gr.176214.114

Kelley, J. L., Peyton, J. T., Fiston-Lavier, A.-S., Teets, N. M., Yee, M.-C., Johnston, J. S., et al. (2014). Compact genome of the Antarctic midge is likely an adaptation to an extreme environment. *Nat. Commun.* 5, 4611. doi:10.1038/ncomms5611

Khan, S., and Vihinen, M. (2010). Performance of protein stability predictors. *Hum. Mutat.* 31, 675–684. doi:10.1002/humu.21242

Kim, I., Miller, C. R., Young, D. L., and Fields, S. (2013). High-throughput analysis of in vivo protein stability. *Mol. Cell. Proteomics* 12, 3370–3378. doi:10.1074/mcp.O113.031708

Kim, J., and Reed, J. L. (2012). RELATCH: relative optimality in metabolic networks explains robust metabolic and regulatory responses to perturbations. *Genome Biol.* 13, R78. doi:10.1186/gb-2012-13-9-r78

Kircher, M., Witten, D. M., Jain, P., O'Roak, B. J., Cooper, G. M., and Shendure, J. (2014). A general framework for estimating the relative pathogenicity of human genetic variants. *Nat. Genet.* 46, 310–315. doi:10.1038/ng.2892

Kosuri, S., Goodman, D. B., Cambray, G., Mutalik, V. K., Gao, Y., Arkin, A. P., et al. (2013). Composability of regulatory sequences controlling transcription and translation in *Escherichia coli. Proc. Natl. Acad. Sci. U.S.A.* 110, 14024–14029. doi:10.1073/pnas.1301301110

Krishnan, V. G., and Westhead, D. R. (2003). A comparative study of machine-learning methods to predict the effects of single nucleotide polymorphisms on protein function. *Bioinformatics* 19, 2199–2209. doi:10.1093/bioinformatics/btg297

Kumar, P., Henikoff, S., and Ng, P. C. (2009). Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nat. Protoc.* 4, 1073–1081. doi:10.1038/nprot.2009.86

Lee, J., Yun, H., Feist, A. M., Palsson, B. Ø, and Lee, S. Y. (2008). Genome-scale reconstruction and in silico analysis of the *Clostridium acetobutylicum* ATCC 824 metabolic network. *Appl. Microbiol. Biotechnol.* 80, 849–862. doi:10.1007/s00253-008-1654-4

Lehner, B. (2013). Genotype to phenotype: lessons from model organisms for human genetics. *Nat. Rev. Genet.* 14, 168–178. doi:10.1038/nrg3404

Lewis, N. E., Nagarajan, H., and Palsson, B. O. (2012). Constraining the metabolic genotype–phenotype relationship using a phylogeny of in silico methods. *Nat. Rev. Microbiol.* 10, 291–305. doi:10.1038/nrmicro2737

Li, B., Krishnan, V. G., Mort, M. E., Xin, F., Kamati, K. K., Cooper, D. N., et al. (2009). Automated inference of molecular mechanisms of disease from amino acid substitutions. *Bioinformatics* 25, 2744–2750. doi:10.1093/bioinformatics/btp528

Li, R., Fan, W., Tian, G., Zhu, H., He, L., Cai, J., et al. (2010). The sequence and de novo assembly of the giant panda genome. *Nature* 463, 311–317. doi:10.1038/nature08696

Liu, J. K., O'Brien, E. J., Lerman, J. A., Zengler, K., Palsson, B. Ø, and Feist, A. M. (2014). Reconstruction and modeling protein translocation and compartmentalization in *Escherichia coli* at the genome-scale. *BMC Syst. Biol.* 8:110. doi:10.1186/s12918-014-0110-6

Long, M., Betrán, E., Thornton, K., and Wang, W. (2003). The origin of new genes: glimpses from the young and old. *Nat. Rev. Genet.* 4, 865–875. doi:10.1038/nrg1204

Machado, D., and Herrgård, M. (2014). Systematic evaluation of methods for integration of transcriptomic data into constraint-based models of metabolism. *PLoS Comput. Biol.* 10:e1003580. doi:10.1371/journal.pcbi.1003580

Machado, D., Rodrigues, L. R., and Rocha, I. (2014). A kinetic model for curcumin production in *Escherichia coli. BioSystems* 125, 16–21. doi:10.1016/j.biosystems.2014.09.001

Masso, M., and Vaisman, I. I. (2010). Knowledge-based computational mutagenesis for predicting the disease potential of human non-synonymous single nucleotide polymorphisms. *J. Theor. Biol.* 266, 560–568. doi:10.1016/j.jtbi.2010.07.026

Melamed, D., Young, D. L., Gamble, C. E., Miller, C. R., and Fields, S. (2013). Deep mutational scanning of an RRM domain of the *Saccharomyces cerevisiae* poly(A)-binding protein. *RNA* 19, 1537–1551. doi:10.1261/rna.040709.113

Milne, C. B., Kim, P.-J., Eddy, J. A., and Price, N. D. (2009). Accomplishments in genome-scale in silico modeling for industrial and medical biotechnology. *Biotechnol. J.* 4, 1653–1670. doi:10.1002/biot.200900234

Monk, J., and Palsson, B. Ø (2014). Predicting microbial growth. *Science* 344, 1448–1449. doi:10.1126/science.1253388

Nakamura, Y., Mori, K., Saitoh, K., Oshima, K., Mekuchi, M., Sugaya, T., et al. (2013). Evolutionary changes of multiple visual pigment genes in the complete genome of Pacific bluefin tuna. *Proc. Natl. Acad. Sci. U.S.A.* 110, 11061–11066. doi:10.1073/pnas.1302051110

Nam, H., Campodonico, M., Bordbar, A., Hyduke, D. R., Kim, S., Zielinski, D. C., et al. (2014). A systems approach to predict oncometabolites via context-specific genome-scale metabolic networks. *PLoS Comput. Biol.* 10:e1003837. doi:10.1371/journal.pcbi.1003837

Ng, P. C., and Henikoff, S. (2001). Predicting deleterious amino acid substitutions. *Genome Res.* 11, 863–874. doi:10.1101/gr.176601

O'Brien, E. J., Lerman, J. A., Chang, R. L., Hyduke, D. R., and Palsson, B. Ø (2013). Genome-scale models of metabolism and gene expression extend and refine growth phenotype prediction. *Mol. Syst. Biol.* 9, 693–693. doi:10.1038/msb.2013.52

Okser, S., Pahikkala, T., Airola, A., Salakoski, T., Ripatti, S., and Aittokallio, T. (2014). Regularized machine learning in the genetic prediction of complex traits. *PLoS Genet.* 10:e1004754. doi:10.1371/journal.pgen.1004754

Orth, J. D., Thiele, I., and Palsson, B. Ø (2010). What is flux balance analysis? *Nat. Biotechnol.* 28, 245–248. doi:10.1038/nbt.1614

Pappalardo, M., and Wass, M. N. (2014). VarMod: modelling the functional effects of non-synonymous variants. *Nucleic Acids Res.* 42, W331–W336. doi:10.1093/nar/gku483

Park, J. H., Lee, K. H., Kim, T. Y., and Lee, S. Y. (2007). Metabolic engineering of *Escherichia coli* for the production of L-valine based on transcriptome analysis and in silico gene knockout simulation. *Proc. Natl. Acad. Sci. U.S.A.* 104, 7797–7802. doi:10.1073/pnas.0702609104

Pegadaraju, V., Nipper, R., Hulke, B., Qi, L., and Schultz, Q. (2013). De novo sequencing of sunflower genome for SNP discovery using RAD (restriction site associated DNA) approach. *BMC Genomics* 14:556. doi:10.1186/1471-2164-14-556

Peskov, K., Mogilevskaya, E., and Demin, O. (2012). Kinetic modelling of central carbon metabolism in *Escherichia coli. FEBS J.* 279, 3374–3385. doi:10.1111/j.1742-4658.2012.08719.x

Price, N. D., Reed, J. L., and Palsson, B. Ø (2004). Genome-scale models of microbial cells: evaluating the consequences of constraints. *Nat. Rev. Microbiol.* 2, 886–897. doi:10.1038/nrmicro1023

Puchałka, J., Oberhardt, M. A., Godinho, M., Bielecka, A., Regenhardt, D., Timmis, K. N., et al. (2008). Genome-scale reconstruction and analysis of the *Pseudomonas putida* KT2440 metabolic network facilitates applications in biotechnology. *PLoS Comput. Biol.* 4:e1000210. doi:10.1371/journal.pcbi.1000210

Ramensky, V. (2002). Human non-synonymous SNPs: server and survey. *Nucleic Acids Res.* 30, 3894–3900. doi:10.1093/nar/gkf493

Reva, B., Antipin, Y., and Sander, C. (2011). Predicting the functional impact of protein mutations: application to cancer genomics. *Nucleic Acids Res.* 39, e118–e118. doi:10.1093/nar/gkr407

Saunders, C. T., and Baker, D. (2002). Evaluation of structural and evolutionary contributions to deleterious mutation prediction. *J. Mol. Biol.* 322, 891–901. doi:10.1016/S0022-2836(02)00813-6

Schwarz, J. M., Cooper, D. N., Schuelke, M., and Seelow, D. (2014). MutationTaster2: mutation prediction for the deep-sequencing age. *Nat. Methods* 11, 361–362. doi:10.1038/nmeth.2890

Segrè, 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

Shihab, H. A., Gough, J., Cooper, D. N., Stenson, P. D., Barker, G. L. A., Edwards, K. J., et al. (2012). Predicting the functional molecular, and phenotypic consequences of amino acid substitutions using hidden Markov models. *Hum. Mutat.* 34, 57–65. doi:10.1002/humu.22225

Shlomi, T., Berkman, O., and Ruppin, E. (2005). Regulatory on/off minimization of metabolic flux changes after genetic perturbations. *Proc. Natl. Acad. Sci. U.S.A.* 102, 7695–7700. doi:10.1073/pnas.0406346102

Skelly, D. A., Merrihew, G. E., Riffle, M., Connelly, C. F., Kerr, E. O., Johansson, M., et al. (2013). Integrative phenomics reveals insight into the structure of phenotypic diversity in budding yeast. *Genome Res.* 23, 1496–1504. doi:10.1101/gr.155762.113

Snitkin, E. S., Dudley, A. M., Janse, D. M., Wong, K., Church, G. M., and Segrè, D. (2008). Model-driven analysis of experimentally determined growth phenotypes for 465 yeast gene deletion mutants under 16 different conditions. *Genome Biol.* 9, R140. doi:10.1186/gb-2008-9-9-r140

Soares-Castro, P., and Santos, P. M. (2013). Towards the description of the genome catalogue of *Pseudomonas sp.* strain M1. *Genome Announc.* 1, e146–e112. doi:10.1128/genomeA.00146-12

Stanford, N. J., Lubitz, T., Smallbone, K., Klipp, E., Mendes, P., and Liebermeister, W. (2013). Systematic construction of kinetic models from genome-scale metabolic networks. *PLoS One* 8:e79195. doi:10.1371/journal.pone.0079195

Stone, E. A. (2005). Physicochemical constraint violation by missense substitutions mediates impairment of protein function and disease severity. *Genome Res.* 15, 978–986. doi:10.1101/gr.3804205

Sunyaev, S. (2001). Prediction of deleterious human alleles. *Hum. Mol. Genet.* 10, 591–597. doi:10.1093/hmg/10.6.591

Szappanos, B., Kovács, K., Szamecz, B., Honti, F., Costanzo, M., Baryshnikova, A., et al. (2011). An integrated approach to characterize genetic interaction networks in yeast metabolism. *Nat. Genet.* 43, 656–662. doi:10.1038/ng.846

Tavtigian, S. V. (2005). Comprehensive statistical study of 452 BRCA1 missense substitutions with classification of eight recurrent substitutions as neutral. *J. Med. Genet.* 43, 295–305. doi:10.1136/jmg.2005.033878

Tepper, N., and Shlomi, T. (2009). Predicting metabolic engineering knockout strategies for chemical production: accounting for competing pathways. *Bioinformatics* 26, 536–543. doi:10.1093/bioinformatics/btp704

Terp, B. N., Cooper, D. N., Christensen, I. T., Jøgensen, F. S., Bross, P., Gregersen, N., et al. (2002). Assessing the relative importance of the biophysical properties of amino acid substitutions associated with human genetic disease. *Hum. Mutat.* 20, 98–109. doi:10.1002/humu.10095

Thiele, I., and Palsson, B. Ø. (2010). A protocol for generating a high-quality genome-scale metabolic reconstruction. *Nat. Protoc.* 5, 93–121. doi:10.1038/nprot.2009.203

Thiele, I., Swainston, N., Fleming, R. M., Hoppe, A., Sahoo, S., Aurich, M. K., et al. (2013). A community-driven global reconstruction of human metabolism. *Nat. Biotechnol.* 31, 419–425. doi:10.1038/nbt.2488

Thomas, P. D., and Kejariwal, A. (2004). Coding single-nucleotide polymorphisms associated with complex vs. Mendelian disease: evolutionary evidence for differences in molecular effects. *Proc. Natl. Acad. Sci. U.S.A.* 101, 15398–15403. doi:10.1073/pnas.0404380101

Thusberg, J., Olatubosun, A., and Vihinen, M. (2011). Performance of mutation pathogenicity prediction methods on missense variants. *Hum. Mutat.* 32, 358–368. doi:10.1002/humu.21445

Tian, J., Wu, N., Guo, X., Guo, J., Zhang, J., and Fan, Y. (2007). Predicting the phenotypic effects of non-synonymous single nucleotide polymorphisms based on support vector machines. *BMC Bioinformatics* 8:450. doi:10.1186/1471-2105-8-450

van Opijnen, T., and Camilli, A. (2013). Transposon insertion sequencing: a new tool for systems-level analysis of microorganisms. *Nat. Rev. Microbiol.* 11, 435–442. doi:10.1038/nrmicro3033

Varma, A., and Palsson, B. Ø (1993). Metabolic capabilities of *Escherichia coli* II. optimal growth patterns. *J. Theor. Biol.* 165, 503–522. doi:10.1006/jtbi.1993.1203

Wang, H. H., Isaacs, F. J., Carr, P. A., Sun, Z. Z., Xu, G., Forest, C. R., et al. (2009). Programming cells by multiplex genome engineering and accelerated evolution. *Nature* 460, 894–898. doi:10.1038/nature08187

Wang, L., Han, X., Zhang, Y., Li, D., Wei, X., Ding, X., et al. (2014). Deep resequencing reveals allelic variation in Sesamum indicum. *BMC Plant Biol.* 14:225. doi:10.1186/s12870-014-0225-3

Wang, M., Zhao, X., Takemoto, K., Xu, H., Li, Y., Akutsu, T., et al. (2012). FunSAV: predicting the functional effect of single amino acid variants using a two-stage random forest model. *PLoS One* 7:e43847. doi:10.1371/journal.pone.0043847

Xu, T., Li, Y., Nostrand, J. D. V., He, Z., and Zhou, J. (2014). Cas9-based tools for targeted genome editing and transcriptional control. *Appl. Environ. Microbiol.* 80, 1544–1552. doi:10.1128/AEM.03786-13

Yamamoto, K., Tamaki, H., Cadillo-Quiroz, H., Imachi, H., Kyrpides, N., Woyke, T., et al. (2014). Complete genome sequence of *Methanoregula formicica* SMSPT a mesophilic hydrogenotrophic methanogen isolated from a methanogenic upflow anaerobic sludge blanket reactor. *Genome Announc.* 2, e870–e814. doi:10.1128/genomeA.00870-14

Yang, H., Krumholz, E. W., Brutinel, E. D., Palani, N. P., Sadowsky, M. J., Odlyzko, A. M., et al. (2014). Genome-scale metabolic network validation of *Shewanella oneidensis* using transposon insertion frequency analysis. *PLoS Comput. Biol.* 10:e1003848. doi:10.1371/journal.pcbi.1003848

Yates, C. M., Filippis, I., Kelley, L. A., and Sternberg, M. J. E. (2014). SuSPect: enhanced prediction of single amino acid variant (SAV) phenotype using network features. *J. Mol. Biol.* 426, 2692–2701. doi:10.1016/j.jmb.2014.04.026

Ye, Z.-Q., Zhao, S.-Q., Gao, G., Liu, X.-Q., Langlois, R. E., Lu, H., et al. (2007). Finding new structural and sequence attributes to predict possible disease association of single amino acid polymorphism (SAP). *Bioinformatics* 23, 1444–1450. doi:10.1093/bioinformatics/btm119

Yue, P., Melamud, E., and Moult, J. (2006). SNPs3D: candidate gene and SNP selection for association studies. *BMC Bioinformatics* 7:166. doi:10.1186/1471-2105-7-166

Zhao, N., Han, J. G., Shyu, C.-R., and Korkin, D. (2014). Determining effects of non-synonymous SNPs on protein-protein interactions using supervised and semi-supervised learning. *PLoS Comput. Biol.* 10:e1003592. doi:10.1371/journal.pcbi.1003592

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.

Received: 25 November 2014; paper pending published: 25 December 2014; accepted: 22 January 2015; published online: 16 February 2015.

Citation: Cardoso JGR, Andersen MR, Herrgård MJ and Sonnenschein N (2015) Analysis of genetic variation and potential applications in genome-scale metabolic modeling. Front. Bioeng. Biotechnol. ***3****:13. doi: 10.3389/fbioe.2015.00013*

This article was submitted to Systems Biology, a section of the journal Frontiers in Bioengineering and Biotechnology.

frontiers
in Cell and Developmental Biology

METHODS
published: 24 March 2015
doi: 10.3389/fcell.2015.00017

RobOKoD: microbial strain design for (over)production of target compounds

Natalie J. Stanford[1, 2]*, **Pierre Millard**[1, 2, 3, 4, 5] and **Neil Swainston**[1, 2]

[1] Manchester Institute of Biotechnology, University of Manchester, Manchester, UK, [2] School of Computer Science, University of Manchester, Manchester, UK, [3] INSA, UPS, INP, LISBP, Université de Toulouse, Toulouse, France, [4] INRA, UMR792, Ingénierie des Systèmes Biologiques et des Procédés, Toulouse, France, [5] Centre National de la Recherche Scientifique, UMR5504, Toulouse, France

Sustainable production of target compounds such as biofuels and high-value chemicals for pharmaceutical, agrochemical, and chemical industries is becoming an increasing priority given their current dependency upon diminishing petrochemical resources. Designing these strains is difficult, with current methods focusing primarily on knocking-out genes, dismissing other vital steps of strain design including the overexpression and dampening of genes. The design predictions from current methods also do not translate well-into successful strains in the laboratory. Here, we introduce RobOKoD (Robust, Overexpression, Knockout and Dampening), a method for predicting strain designs for overproduction of targets. The method uses flux variability analysis to profile each reaction within the system under differing production percentages of target-compound and biomass. Using these profiles, reactions are identified as potential knockout, overexpression, or dampening targets. The identified reactions are ranked according to their suitability, providing flexibility in strain design for users. The software was tested by designing a butanol-producing *Escherichia coli* strain, and was compared against the popular OptKnock and RobustKnock methods. RobOKoD shows favorable design predictions, when predictions from these methods are compared to a successful butanol-producing experimentally-validated strain. Overall RobOKoD provides users with rankings of predicted beneficial genetic interventions with which to support optimized strain design.

Keywords: synthetic biology, systems biology, metabolic engineering, strain design, constraint-based modeling

Edited by:
Markus J. Herrgard,
Technical University of Denmark,
Denmark

Reviewed by:
Monika Heiner,
Brandenburg University of Technology
Cottbus-Senftenberg, Germany
Osbaldo Resendis-Antonio,
Instituto Nacional de Medicina
Genomica, Mexico

Correspondence:
Natalie J. Stanford,
School of Computer Science,
University of Manchester, Oxford
Road, Manchester M13 9PL, UK
natalie.stanford@manchester.ac.uk

Specialty section:
This article was submitted to Systems Biology, a section of the journal Frontiers in Cell and Developmental Biology

Received: *03 November 2014*
Accepted: *25 February 2015*
Published: *24 March 2015*

Citation:
Stanford NJ, Millard P and Swainston N (2015) RobOKoD: microbial strain design for (over)production of target compounds.
Front. Cell Dev. Biol. 3:17.
doi: 10.3389/fcell.2015.00017

Introduction

The sustainable production of target compounds such as biofuels and high-value chemicals for pharmaceutical, agrochemical, and chemical industries is becoming an increasing priority given their current dependency upon diminishing petrochemical resources. The challenge of producing such compounds from microbial cells straddles both systems and synthetic biology. The development of microbial cell factories first requires a comprehensive understanding of host cell metabolic functions through metabolic model construction, and subsequent *in silico* experimentation, using systems biology methods. This *in silico* experimentation can suggest host cell manipulations that can be applied *in vitro* using synthetic biology techniques, leading to increased production of the target compound (Koide et al., 2009).

Target producing microbial strains are typically designed using combinations of gene manipulations. These manipulations include gene additions (often recombinant genes from other organisms) and removal of genes via knockouts. Furthermore, over-expression or inhibition of host genes can either increase or dampen metabolic flux through the reactions that their expressed proteins catalyze. Successful application of such strategies can be used to overproduce host-native targets (Ng et al., 2012; Li et al., 2014) or produce non-host-native targets (Atsumi et al., 2009; Angermayr et al., 2014; Yuan et al., 2014). Identifying successful gene manipulation combinations has traditionally relied on static network inspection, and experimental trial and error to test the strategies (Varman et al., 2011). This approach is not optimal as it limits the amount of network information that can be used, discounts metabolic complexity, and therefore prevents predictions of less intuitive metabolic modifications (Kitano, 2002).

Through modeling approaches, strain predictions can be improved by taking into account full metabolic complexity during the design phase. Designed strains can also be screened *in silico* before they are engineered and tested in the laboratory. The process involves iterative application of the following steps: (i) characterization of the host metabolic network; (ii) identification of gene additions to bridge native metabolism to the target; (iii) optimization of the modified metabolic network through gene addition, deletion, overexpression or dampening; (iv) trialing successful predictions in the laboratory. This process affords the potential to develop successful strains more cost effectively, and time efficiently. This work focuses on step (iii), which involves elements of network characterization in order to identify suitable optimization strategies.

To characterize the metabolic network, genome-scale models (GEMs) can be used in conjunction with constraint-based techniques. GEMs are computer-analyzable, structured knowledge bases of genes, proteins, and metabolites present within a given organism (Thiele and Palsson, 2010). GEMs therefore encode the complexity of host cell metabolism and are available for an increasingly large number of organisms (Büchel et al., 2013). Constraint based techniques, including flux balance analysis (FBA) and flux variability analysis (FVA), provide quantitative predictions of cellular behavior such as metabolic flux patterns and cellular growth rates. These are computed by applying constraints, which can be assigned from experimentally measured nutrient uptake rates (Orth et al., 2010) and intracellular fluxes (Sauer, 2006), or inferred through interpretation of gene expression data (Lee et al., 2012). These predictions provide insights into the metabolic pathways active under different growth conditions (Liao et al., 2011), gene essentiality (Joyce and Palsson, 2008; Dobson et al., 2010; Heavner et al., 2012), and as a result, the fitness optimality of a given strain (Harcombe et al., 2013). More detailed introductions to these techniques can be found in **Boxes 1**, **2**.

Optimization of microbial strains is complex, requiring a balance between target production and cell viability (Lo et al., 2013). This makes the problem a multi-objective optimization problem, whereby metabolic flux of cellular growth and target production must be considered simultaneously. Successful optimization strategies therefore include gene modifications (knockouts, over-expression, dampening) which re-route flux toward the target product whilst minimizing the effect on flux toward synthesis of metabolites required for cellular maintenance.

Amongst the more prominent methods used for identifying knockout targets are OptKnock (Burgard et al., 2003) and RobustKnock (Tepper and Shlomi, 2010). OptKnock aims to optimize the maximum flux toward the target product whilst retaining cell viability, using up to five reactions knockouts to generate the strain solution. The method does not take into consideration flux variability, and therefore whilst there may be a reasonable maximal flux yield toward to target product, it is possible that the minimal flux toward the target product could be zero. RobustKnock was developed to improve on this shortcoming, by optimizing the minimal flux toward the target product, again by applying up to five reaction knockouts. Limitations of these methods include the prediction of only a single gene knockout strategy, and also no consideration of over-expression or dampening targets, which are key aspects of successful strain design (Dellomonaco et al., 2011). A complementary method, optGene (later updated to optFlux (Rocha et al., 2010)), can be used for overexpression analysis. Flux Variability Analysis has been used in a number of studies for identifying overexpression targets (Choi et al., 2010; Park et al., 2012), as well as more comprehensive strategies (Pharkya and Maranas, 2006; Feist et al., 2010), although these have not been extensively used. Elementary modes have also been used to identify suitable knockout targets (Ballerstein et al., 2012; von Kamp and Klamt, 2014).

To integrate the requirements of predicting both knockouts and over-/under-expressions, we introduce RobOKoD (Robust Overexpression, Knockout and Dampening). RobOKoD takes into consideration metabolite centrality and flux variability in order to comprehensively identify potential knockouts and gene over-/under-expressions, ranked by significance, and follow the schematic presented in **Figure 1**. This ranking is a strength, as it allows for further, manual analysis of the system to be used for strain design.

The performance of RobOKoD was tested against that of OptKnock and RobustKnock in their ability to predict an engineering strategy for production of butanol from *Escherichia coli* using the reverse β-oxidation cycle. The predictions were validated against a successful, experimentally-validated butanol producing strain developed by Dellomonaco et al. (2011).

Materials and Methods

Escherichia coli model

The model used in this study is a derivation of a core metabolism model derived from the iAF1260 reconstruction of *E. coli* metabolism proposed by Feist et al. (2007). The core metabolism model of 95 native reactions was modified to include the β-oxidation pathway—a total of eight genes catalyzing 30 additional reactions—to produce the model iNS142 (see **Table 1**). This model contains 142 genes, 125 reactions, and 93 metabolites (**Figure 2**). The model is available in Supplementary Folder 1 in SBML format (Hucka et al., 2003).

BOX 1

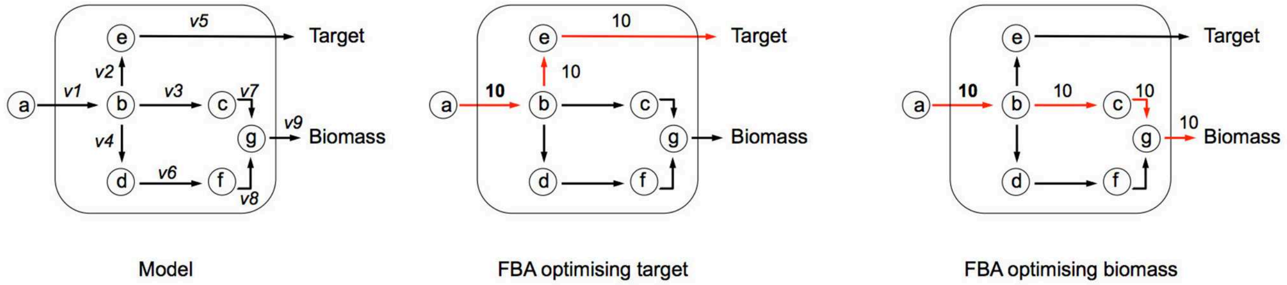
Flux Balance Analysis (FBA) allows the computation of fluxes, and cellular growth, by using a set of constraints. FBA uses the stoichiometric matrix (*S*), which is a matrix consisting of rows of metabolites (*m*), and columns of reactions (*n*). An example based on the toy network in **Figure B1a** can be seen in **Table B1a**. The matrix is usually sparse and filled with positive (negative) coefficients for metabolites produced (consumed) by a reaction. Linear programming is used to compute feasible fluxes (*v*) through the network ensuring that a steady state is satisfied (Equation i), subject to a set of constraints (Equation ii) and optimizing (*Z*) a specific function (Equation iii, where *c* is a vector of weights, typically a vector of zeros with biomass production set to 1). The minimum solutions of Equation (i) are elementary modes, which are minimal sets of enzymes that can operate at steady state, also known as minimal functional units (de Figueiredo et al., 2009). If Equation (i) cannot be satisfied, then FBA cannot be computed on the system.

$$Sv = 0 \quad (i)$$

$$lb_i \leq v_i \geq ub_i, i = 1, \ldots, n \quad (ii)$$

$$Z = c^T v \quad (iii)$$

In the example network below (**Figure B1a**), *c* is given as an uptake rate of 10 units of metabolite **a**. In the center network $Z =$ Target, and in the right-hand network $Z =$ Biomass. Reaction bounds are all assigned as $lb_i = 0$, $ub_i = 1000$. Meaning that each reaction through the network is irreversible. Computing FBA for $Z =$ Target we get 10 units of flux flowing through *v2* and *v3*, producing $v_{\_Target} = 10$ units. For $Z =$ Biomass we get 10 units of flux flowing through *v3*, *v7*, and *v9*, producing $v_{\_Biomass} = 10$ units.



Model

FBA optimising target

FBA optimising biomass

FIGURE B1a | Illustrating FBA for independent optimisation of target and biomass.

TABLE B1a | Stoichiometric matrix (S).

	v1	***v2***	***v3***	***v4***	***v5***	***v6***	***v7***	***v8***	***v9***
a	−1	0	0	0	0	0	0	0	0
b	+1	−1	−1	−1	0	0	0	0	0
c	0	0	+1	0	0	0	−1	0	0
d	0	0	0	+1	0	−1	0	0	0
e	0	+1	0	0	−1	0	0	0	0
f	0	0	0	0	0	+1	0	−1	0
g	0	0	0	0	0	0	+1	+1	−1
bio.	0	0	0	0	0	0	0	0	+1
tar.	0	0	0	0	+1	0	0	0	0

RobOKoD

The RobOKoD method is based on the two following assumptions:

(1) To achieve target production, carbon transfer within the network has to be oriented toward pathways that favor the target. Therefore, changes within the network should aim to reduce carbon loss to peripheral pathways.

(2) Flux variability of each reaction will differ depending on whether the reaction is important for growth, generating the desired product, both, or neither. Therefore, the functionality of each reaction can be inferred by analyzing its variability.

A simplified schematic of the method based on these two assumptions can be seen in **Figure 1** and additional details are given in the next sections. First, a metabolite consumption test (MCT) is applied which computes whether a given metabolite in the target production pathway demonstrates flux loss to biomass production. If flux loss is identified, all reactions that consume that metabolite are flagged as potentially favored targets. Second, flux variability analysis profiling (FVAp) is performed to determine the flux variability of each reaction, at increments of maximum biomass flux and then at increments of maximum target product flux. The profiles of each reaction are used to calculate a score from which the importance of each reaction for growth and target production can be estimated. Finally, MCT and FVAp results are combined to rank potential modifications.

BOX 2

Flux Variability Analysis (FVA). Box 1 showed an example of FBA, where a single set of fluxes was identified, which can maximize biomass production (Z). It can be seen in the central network of **Figure B2a**, that this set of fluxes was just one of two possible solutions that could be selected to maximize Z—route A and route B. FVA allows us to garner this additional information by identifying the minimum and maximum flux that each reaction can carry (Equation i). FVA can be implemented at the optimal state whereby $y = 1$ (Equation ii), subject to flux constraints for each reaction (Equation iii) as demonstrated in the right-hand network in **Figure B2a** (Gudmundsson and Thiele, 2010). Here the main information identified is which reactions are interchangeable. It is also common to compute FVA under suboptimal conditions (i.e., $y = 0.95$ as used in RobOKoD), which introduces a small amount of flexibility in the system and reduces the chances of optimal pathways being unrealistic when compared *in vivo*.

$$v_{\max}/v_{\min} v_i \qquad (i)$$

$$\gamma Z_0 \leq c^{\mathrm{T}} v \qquad (ii)$$

$$v_{lb} \leq v \leq v_{ub} \qquad (iii)$$

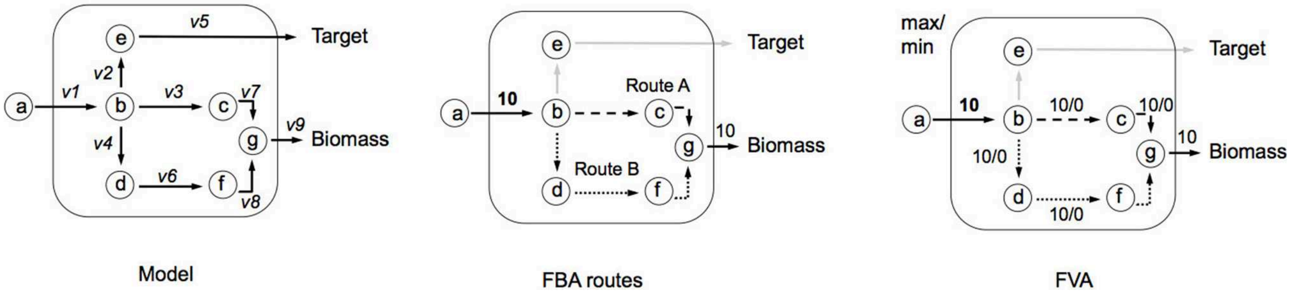
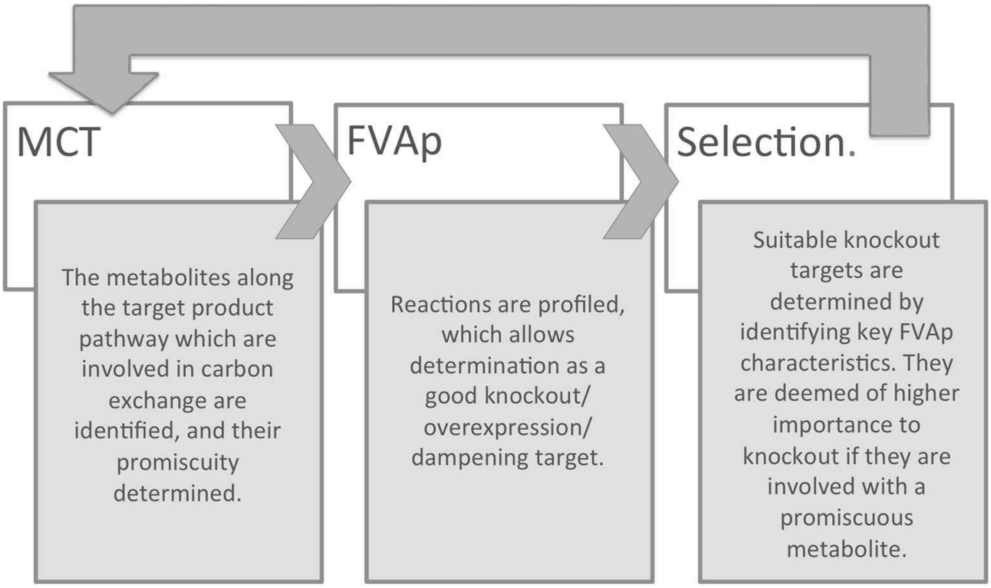


FIGURE B2a | Illustrating implementation of FVA and how it can be used to identify alternative flux optima.



The metabolites along the target product pathway which are involved in carbon exchange are identified, and their promiscuity determined.

Reactions are profiled, which allows determination as a good knockout/ overexpression/ dampening target.

Suitable knockout targets are determined by identifying key FVAp characteristics. They are deemed of higher importance to knockout if they are involved with a promiscuous metabolite.

FIGURE 1 | Workflow of RobOKoD, illustrating the iterative application of the methods MCT and FVAp.

Modifications can consist of (i) gene deletions; (ii) changes of environmental conditions; (iii) gene over-expressions; and (iv) gene dampenings.

This strategy ensures that reactions that are vital for either growth or target product production, or those that produce key metabolites, are not selected as potential knockouts. Conversely, reactions that (i) significantly divert carbon away from target production; and (ii) consume a metabolite known to promote flux loss from target production; are selected preferentially. Once the first knockout is predicted, the model is modified to block this reaction, and the same selection process is used to select the second reaction to delete. This method can be applied iteratively to predict a number of modifications that should enhance target production whilst maintaining growth.

TABLE 1 | Reactions and genes added to the core iAF1260 model to implement the β-oxidation cycle.

Reaction	Gene(s)	EC
Thiolase	*fadA, fadI*	2.3.1.16
Hydroxyacyl-CoA dehydrogenase	*fadB, fadJ*	1.1.1.35
Enoyl-CoA hydratase	*fadB, fadJ*	4.2.1.17
Enoyl-CoA reductase	*fadE*	1.3.8.1
Alcohol/acetaldehyde dehydrogenase	*frmA, adhP, adhE*	1.1.1.1

All code was developed in Matlab to maintain compatibility with the COBRA Toolbox (Schellenberger et al., 2011), and is available in Supplementary Folder 1.

Metabolite Consumption Test (MCT)

Metabolite Consumption Test (MCT) identifies metabolites within the optimal target production pathway that are also consumed to produce biomass. The MCT score is given in a two-step process. First *flux change* (X_m) per metabolite (m) is calculated, then an MCT-value of 1 is given to all reactions that consume metabolites, denoted by a negative X_m. X_m is calculated according to Equation (1). For each metabolite that is featured in the optimal target producing pathway, for the example network in **Figure 3**, that would be metabolites **a**, **b**, **e**, all producing and consuming reactions are identified. Then per identified reaction, a unitary constant c is calculated which identifies the reaction as a producer (+1) or consumer (−1) of the metabolite during biomass production, thereby indicating whether there is a potential flux loss or gain from that reaction. Each reaction is then weighted (w) according to whether it is vital for both target and biomass (0); or potentially used (1), or not used (0) for biomass production. v is the maximum flux through the reaction during biomass production. All reactions that consume a metabolite m with a negative X_m-value are flagged with a 1 in the corresponding column (see MCT column in **Table 2**).

$$X_m = \sum_{i=1}^{n} c^{(i)} \cdot w^{(i)} \cdot v_{\max}^{(i)} \tag{1}$$

FVA Reaction Profile (FVAp)

Prior to FVAp, FBA is applied to predict the maximal theoretical yield of both biomass (y_{bm}) and target product (y_{target}). FVAp is then performed which computes the flux variability of each reaction: (1) at different percentage (0–100%) of y_{bm} whilst optimizing target product; and (2) at different percentage (0–100%) of y_{target}, whilst optimizing biomass. By computing FVAp the flux capacity of each reaction is profiled over a range of target constraints. The key areas of interest are the extremes of target production, and biomass production. It can be seen in **Figure 5** that the first and last quartile of the x axis for all examples holds the key information from which beneficial genetic interventions can be inferred.

Knockout Scoring

Knockouts were selected by computing a knockout ranking score. The ranking score is calculated for each reaction using FVAp at different percentage (0–100%) of y_{bm} whilst optimizing target product (red shaded area). Let us denote with $(v_{\max})^{target}|_p$ and $(v_{\min})^{target}|_p$ the maximal and minimal flux, respectively of reaction i obtained through FVAp when requiring a percentage p of y_{bm} to be produced while maximizing for product. Likewise let the maximal and minimal flux of reaction i obtained through FVAp when requiring a percentage p of y_{target} to be produced while maximizing for biomass be defined as $(v_{\max})^{biomass}|p$ and $(v_{\min})^{biomass}|p$, respectively. It must be noted that the percentage p refers to either biomass or target product production requirement depending on the objective function.

A suitable knockout target displays the key characteristics shown in **Figure 5A**, where the first quartile of x axis 0-25% of y_{bm} (red shaded area) carries a lower $v^{(i)}{}_{max|target}$, than 75-100% of y_{bm}, which shows that the reaction is required to carry a higher flux to sustain optimal biomass production. This characteristic is captured in Equation (2) (biomass reaction activation). A reduced variability in the fourth quartile also demonstrates a stronger constraint on the flux to produce y_{bm}, this is captured in Equation (3) (product variability area). The final knockout scoring R^i_{KOr} for each reaction was computed according to Equation (4), which takes into account the features of both the biomass reaction activation and product variability area.

Biomass reaction activation:

$$\sum_{p_1=75\%}^{100\%} \left(v_{\max}^{(i)}\right)^{target} |p_1 - \sum_{p_2=0\%}^{25\%} \left(v_{\max}^{(i)}\right)^{target} |p_2 \tag{2}$$

Product variability area:

$$\sum_{p=75\%}^{100\%} \left(v_{\max}^{(i)}\right)^{target} |p - \left(v_{\min}^{(i)}\right)^{target} |p \tag{3}$$

$$R^i_{KOr} = \frac{biomass\ reaction\ activation}{product\ variability\ area} \tag{4}$$

Reactions that obtain a high R^i_{KOr}, are identified as a putative target for knocking out providing it is not a lethal target for the cell. Identified target reactions for knocking out are first ordered by R^i_{KOr}, before secondary sorting by MCT flags. An example of this sorting can be seen in **Table 2** based on the toy network presented in **Figures 3**, **4**.

Over-Expression Ranking

The characteristics of a strong over-expression target can be seen in the lower quartile of x axis in **Figure 5B**, where at 0-25% of y_{bm} (red shaded area) $v^{(i)}{}_{\mathrm{min|target}}$ has a higher flux capacity than 75-100% of y_{target} (blue shaded area), $v^{(i)}{}_{\mathrm{min|biomass}}$ (target extra flux, see Equation 5). A lower variability is also desirable for optimizing target subject to 0-25% of y_{bm} (target variability, Equation 6) as it ensures that the minimum flux the reaction

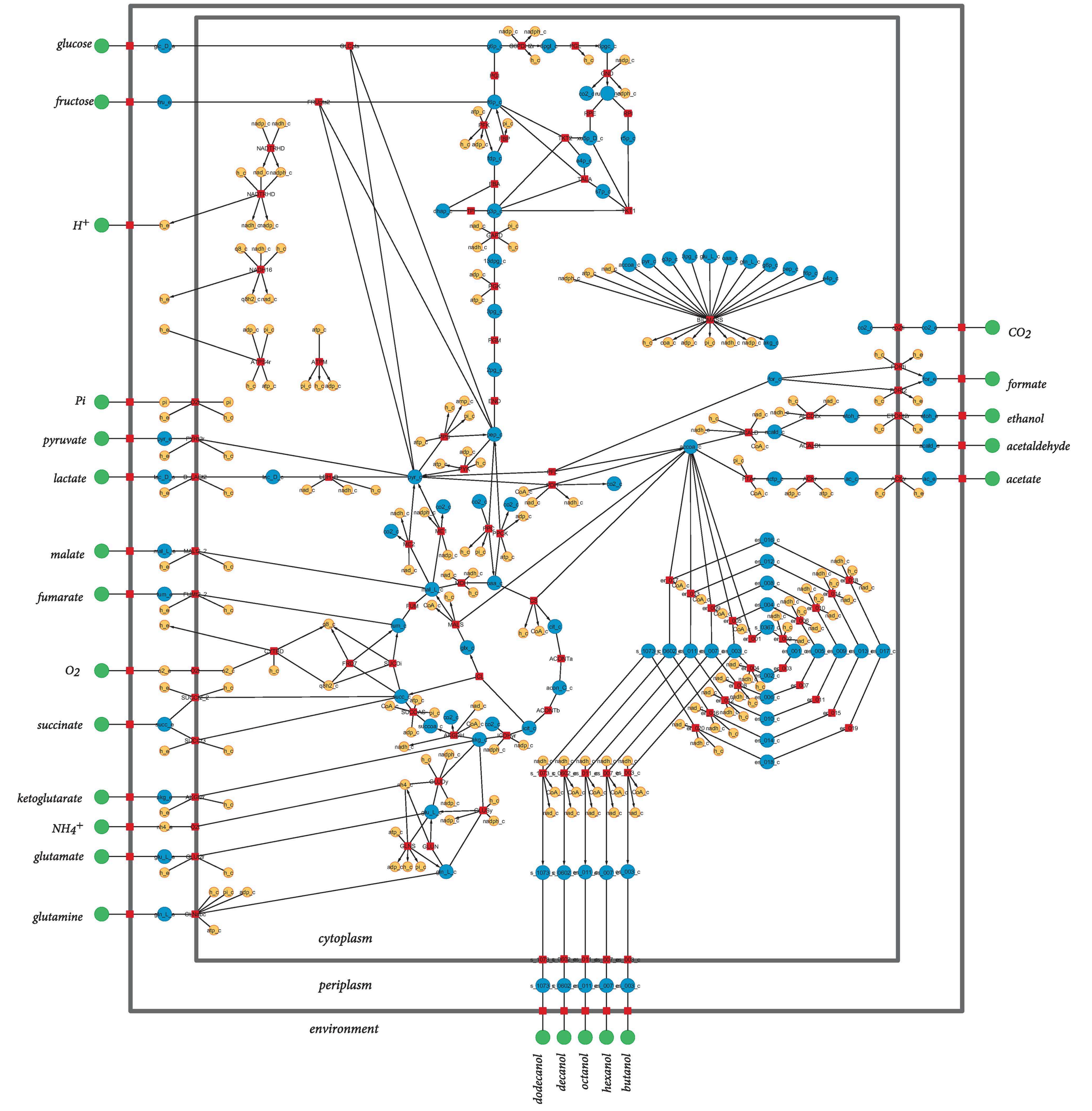


FIGURE 2 | Graphical representation of the metabolic network of *Escherichia coli* included in iNS142. Red squares represent reactions, and green, blue, and orange circles represent extracellular metabolites, intracellular metabolites involved in carbon transfers, and intracellular metabolites not involved in carbon transfers, respectively. Directed arcs show irreversible reactions, whereas undirected arcs show reversible reactions. Water is not shown for clarity of the layout.

can carry is close to optimum. The final ranking (R^i_{OEx}) is determined using Equation (7), where reactions with the highest R^i_{Oex} are the most likely over-expression targets. An example of a weaker over-expression target (corresponding to a lower R^i_{OEx}) is shown in **Figure 5C**, which illustrates an over-expression that will increase flux to both target *and* biomass. Negative R^i_{OEx} represent potential dampening targets (see **Figure 5D**), which display the opposite characteristics.

Target extra flux:

$$\sum_{p_1=0\%}^{25\%} \left(v^{(i)}_{\max}\right)^{target} |p_1 - \sum_{p_2=75\%}^{100\%} \left(v^{(i)}_{max}\right)^{BM} |p_2 \qquad (5)$$

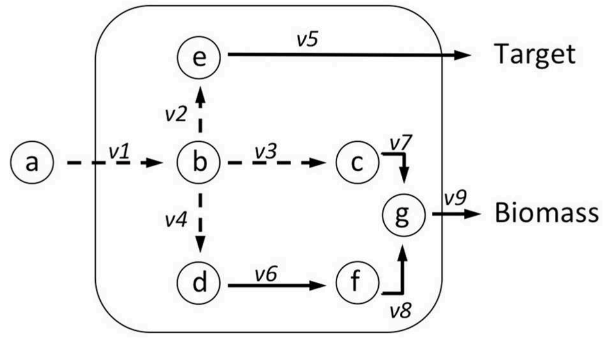


FIGURE 3 | Metabolite Consumption Test (MCT) identifies metabolites that are in the optimal target production pathway. The test has two parts, first a *flux change* (X_m) score is computed using Equation (1). Taking metabolite **b** as an example: *v1* produces **b** but is needed for both target and biomass production so weight (w_1) = 0; *v2* consumes **b** but is needed for producing the target so $w_2 = 0$; *v3* consumes **b**, so $w_3 = 1$; *v4* consumes **b**, so $w_4 = 1$. These values are multiplied by the absolute value of maximum flux calculated using FVA ($v^i_{\max}$), and by a constant (c) = ±1 according to whether the reaction produces or consumes the metabolite. Where $X_m < 0$ MCT = 1, where $X_m \geq 0$ MCT = 0. Reactions identified as suitable knockout targets using RobOKoD are sorted firstly by R^i_{KOr} and secondly by their MCT flag. This means that reactions with an equal R^i_{KOr} can be differentiated by a secondary sorting against whether they directly consume a metabolite that is important for the target production (see **Table 2**).

TABLE 2 | Using the toy network presented in Figures 3, 4 we computed the MCT score and R^i_{KOr} of the intracellular reactions.

Flux	MCT score	R^i_{KOr}
v3	1	0.8523
v4	1	0.8523
v6	0	0.8523
v7	0	0.8523
v8	0	0.8523
v2	0	0

v3, v4, v6, v7, and v8 all have the same FVAp profiles and therefore R^i_{KOr} scores. Of the top ranking reactions within this network v3 and v4 consume a metabolite that is important for target production. These reactions are then sorted as a higher priority within the equally ranked reactions to select as a knockout target.

Target variability:

$$\sum_{p=0\%}^{25\%} \left(v^{(i)}_{\max}\right)^{target} |p - \left(v^{(i)}_{\min}\right)^{target} |p \qquad (6)$$

$$R^i_{OEx} = \frac{target\ extra\ flux}{target\ variability} \qquad (7)$$

OptKnock and RobustKnock

The OptKnock algorithm (Burgard et al., 2003) is available in the COBRA Toolbox for Matlab, and RobustKnock algorithm is available as a Matlab script from the original paper (Tepper and Shlomi, 2010). Both are repackaged in Supplementary File 1 allowing for reproduction of the following results.

Results

As a case study, RobOKoD was applied to design an *E. coli* strain with a reverse β-oxidation cycle for butanol production. These results can be recreated by unzipping the code in Supplementary File 1, and running the test script *iNS142_butanol.m* in Matlab [requires the COBRA Toolbox (Schellenberger et al., 2011), and if RobustKnock is to be tested, the Tomlab solver (Tomlab Optimization Inc., Västerås, Sweden)]. This test script runs RobOKoD over a maximum of five iterations of knockout scoring, implementing the highest scoring knockout, generating a results document and reaction FVA profile plots for each iteration in the directory *iNS142_butanol_results*, and outputting an updated SBML model in which the knockouts have been implemented. It subsequently runs over-expression ranking, again generating output in the *iNS142_butanol_results* directory. OptKnock and RobustKnock are then run in order to compare predictions from each method. Knockout scoring, over-expression rankings, and FVA profiles for all relevant reactions (such as those illustrated in **Figure 3**) can then be inspected manually.

MCT allows the identification of reactions which consume metabolites present in the optimal target production pathway that demonstrate flux loss toward biomass. These reactions are flagged in the listing of potential knockouts with a value of 1, allowing these reactions to be identified preferentially, out a set of reactions with the same knockout score. In this network, pyruvate was identified as a key metabolite where flux loss to biomass production could occur, 11 reactions were then identified that consume pyruvate.

FVA profiles representative of the different situations commonly encountered are shown in **Figure 5**. Knockout targets (**Figure 5A**) are identified based on fixed biomass optimal target FVAp (red profile). As the percentage of fixed biomass increases, the flux through the reaction increases to accommodate a higher biomass requirement, and the variability of the flux narrows. Strong overexpression targets (**Figure 5B**) show the opposite behavior of knockouts, whereby the flux through the reaction reduces as the percentage of fixed target is reduced as biomass is optimized (blue profile). Weak overexpression targets (**Figure 5C**) show similar characteristics, but are not required to carry a flux for the target to be optimized. Dampening targets (**Figure 5D**) are characterized by their ability to carry higher flux through a reactions at low percentage of fixed target with optimized biomass, than at both a high percent of fixed target and optimized biomass, and a low percent of fixed biomass and optimized target.

It is noted that some reactions obtain identical scores, hence their deletion are predicted to have the same impact on the system. This is for instance the case for two consecutive reactions of an unbranched, linear pathway. More generally, this is observed for the subsets of reactions that carry perfectly correlated fluxes (Heiner, 2009; Feist et al., 2010). A feature of RobOKoD is therefore its ability to identify such subsets of reactions. The corresponding knockouts are expected to result in a similar phenotype, hence the modification to perform for such subsets of reactions should be evaluated in the light of technical considerations. The most practical modifications should be selected,

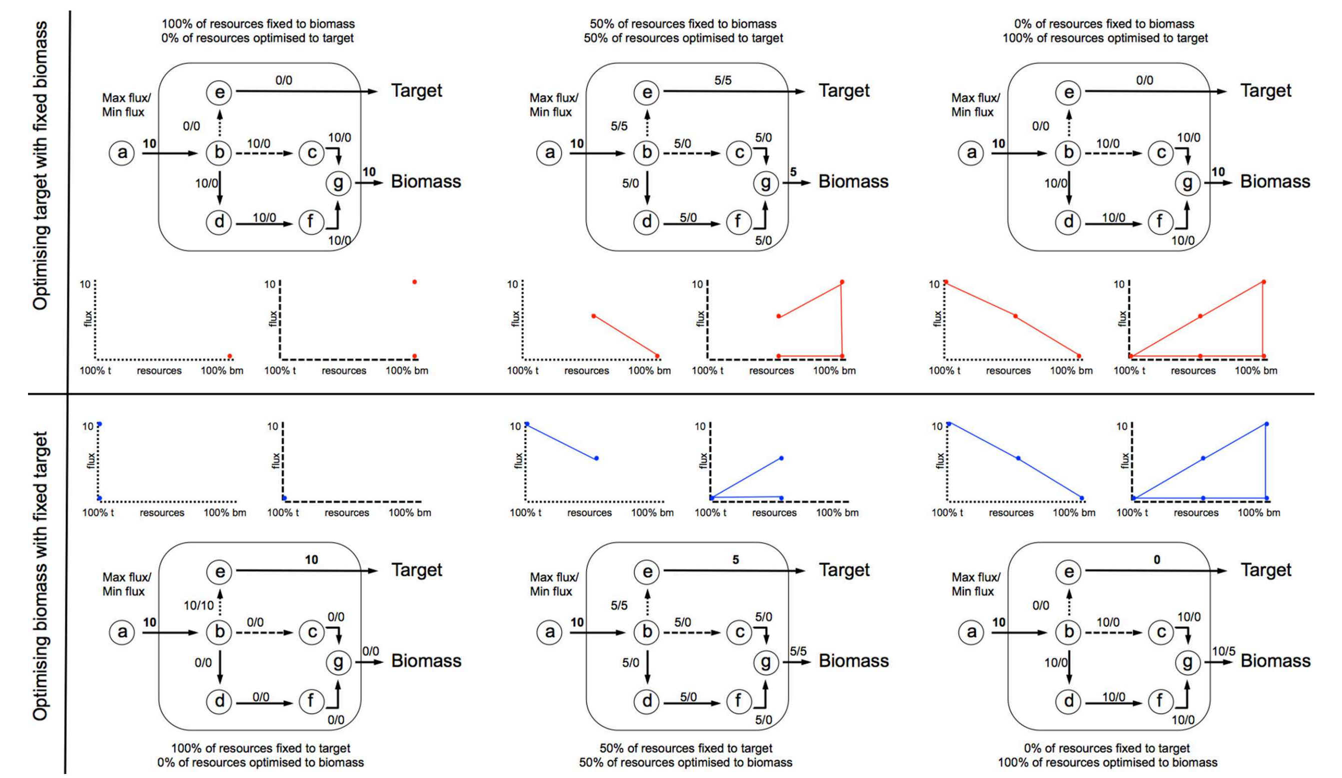

Optimising target with fixed biomass
100% of resources fixed to biomass
0% of resources optimised to target
50% of resources fixed to biomass
50% of resources optimised to target
0% of resources fixed to biomass
100% of resources optimised to target
Max flux/
Min flux
Target
Biomass
a
b
c
d
e
f
g
10
0/0
10/0
5/5
5/0
5
flux
100% t
resources
100% bm
Optimising biomass with fixed target
10/10
10/5
5/5
100% of resources fixed to target
0% of resources optimised to biomass
50% of resources fixed to target
50% of resources optimised to biomass
0% of resources fixed to target
100% of resources optimised to biomass

FIGURE 4 | FVA is performed to compute the flux variability of each reaction: (1) at different percentage (0–100%) of y_{bm} whilst optimizing target product; and (2) at different percentage (0–100%) of y_{target}, whilst optimizing biomass. Each iteration develops a profile of $v_{min}^{(i)}$ and $v_{max}^{(i)}$ across the range of flux space. In this example an input flux of 10 units through v_1 is fixed, and the network optimized to situation 1 (top profile), or 2 (bottom profile).

whilst the resulting strain should still be amongst the optimal producers.

For comparison purpose, the well-established algorithms OptKnock and RobustKnock were applied on the same model to predict the optimal strain for butanol production. For each method, the maximum number of modifications was fixed to five, since constructing such a strain can still be managed experimentally. The optimal producer strains predicted by each method are listed in **Table 3** and are compared to the most efficient producer strain which has been experimentally validated (Dellomonaco et al., 2011). OptKnock and RobustKnock predicted strains that were theoretically unable to produce butanol during growth, and in the case of OptKnock, not viable for growth.

Table 4 compares the functionality modifications of the predicted *in silico* cells, and the experimental strain. It appears that RobOKoD automatically captures most of the functional modifications experimentally carried out. In particular, it predicted that fermentation pathways (*pfl, ldhA*) should be knocked out to avoid diversion of carbon and reduced cofactors toward by-products of poor interest. Moreover, by highlighting the competing interests of oxygen uptake pathway between the production of biomass and butanol, RobOKoD was able to indicate an anoxic condition change, similar to the experimental strain which knocked-out fumarate reductase and was grown under microaerobic conditions.

In addition to the knockout predictions, RobOKoD was also able to predict over-expression and dampening targets. It predicted that enzymes catalyzing the reactions associated with the reverse β-oxidation cycle should be over-expressed, consistent with the experimental strain where the activity of transcriptional inhibitors of this pathway are dampened (*fadR*, *atoC(c)*, *crp**, and Δ*arcA strains*). Moreover, RobOKoD also predicts that a number of transport reactions (or rather *genes* encoding the relevant transport proteins) should be dampened, hence providing additional modifications that could enhance butanol production. These dampening predictions, less intuitive, were not carried out in the experimental strain and have not been experimentally verified.

Table 5 compares the molar production of butanol per mole of glucose uptake, when the objective of the cell is to optimize biomass. It shows that RobOKoD predicted the most successful butanol strain design, with molar ratio values similar to that achieved in the experimental strain. Neither OptKnock or RobustKnock predicted successful strains, and in the case of OptKnock, the strain was predicted to be no longer viable.

The strain predicted by RobOKoD was developed iteratively by automatically knocking out the highest ranked suggested

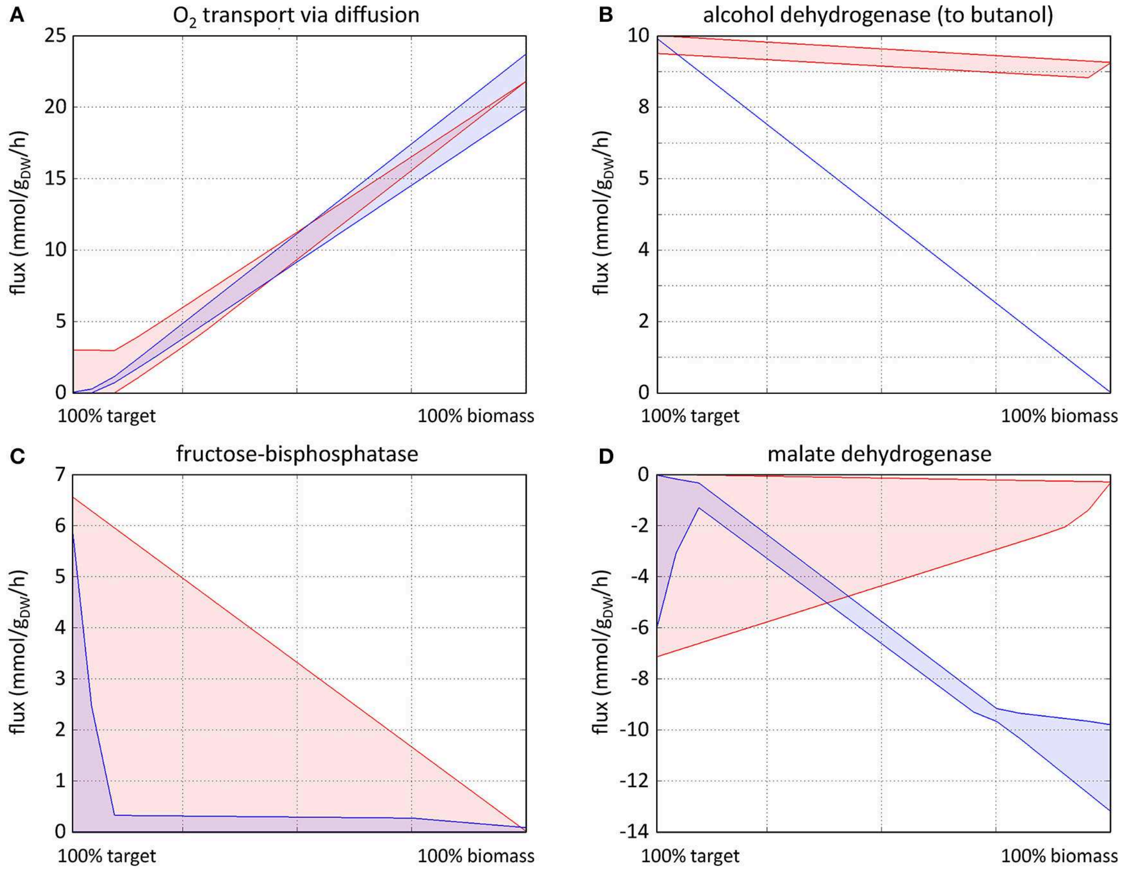


FIGURE 5 | Typical FVA profiles characteristic of knockout targets (A), strong overexpression targets (B), weak overexpression targets (C), and dampening targets (D). The red profiles show FVAp of each reaction at different percentages of (0–100%) of y_{bm} whilst optimizing target product. The blue profiles show FVAp at different percentages of (0–100%) of y_{target} whilst optimizing biomass. Knockout targets **(A)** are identified using (75–100%) of y_{bm} (corresponding to the fourth quartile of *x* axis) with target optimization: where $v^{(i)}_{max|target}$ increases as y_{bm} increases, coupled with a reduced variability between $v^{(i)}_{max|target}$ and $v^{(i)}_{min|target}$. Strong overexpression targets **(B)** are identified using (0–25%) of y_{bm} optimizing target, and (75–100%) of y_{target} optimizing biomass (corresponding to the first quartile of *x* axis), where $v^{(i)}_{max|target}$ (red) has a higher flux carrying capacity than $v^{(i)}_{max|biomass}$ (blue), again with reduced variability between $v^{(i)}_{max|target}$ and $v^{(i)}_{min|target}$. Weak overexpression targets **(C)** show similar characteristics, with a smaller difference between $v^{(i)}_{min|target}$ and $v^{(i)}_{min|biomass}$ and a larger variability between $v^{(i)}_{min|target}$ and $v^{(i)}_{min|target}$. Profiles of dampening targets **(D)** are the reverse of overexpression targets.

TABLE 3 | Gene modifications, based on the reactions predicted by the three computational methods, and their comparison with those successfully applied experimentally (Dellomonaco et al., 2011).

Method	Gene modifications [Δ*gene*$_{(reaction)}$]
OptKnock	Δ*eutE*$_{(ACALD)}$ Δ*nuoH*$_{(NADH16)}$ Δ*amtB*$_{(NH4t)}$ Δ*pflA*$_{(PFL)}$ Δ*pitB*$_{(PIt2r)}$
RobustKnock	Δ*lldP*$_{(D\_LACt2)}$ Δ*focA*$_{(FORti)}$ Δ*pgi*$_{(PGI)}$ Δ*satP*$_{(SUCCt2\_2)}$ Δ*sucD*$_{(SUCOAS)}$
RobOKoD	Anoxic conditions$_{(O2t)}$, Δ*pflA*$_{(PFL)}$, Δ*eutE*$_{(ACALD)}$, Δ*dld*$_{(LDH\_D)}$, *fadA*+, *yqeF*+
Experimental	RB02(*fadR atoC(c) crp** Δ*arcA* Δ*adhE* Δ*pta* Δ*frdA*) Δ*yqhD* Δ*eutE yqeF*+ *fucO*+

TABLE 4 | Functional similarities captured in the gene manipulations predicted by each method.

Gene	Function	OptKnock	RobustKnock	RobOKoD
Δ*adhE*	Alcohol/acetaldehyde dehydrogenase			✓
Δ*pta*	Phosphotransacetylase			✓
Δ*frdA*	Fumarate reductase (respiration)		✓	
Δ*yqhD*	Alcohol dehydrogenase			✓
Δ*eutE*	Acetaldehyde dehydrogenase	✓		✓

knockout target, that also was flagged by MCT as a potential route for flux loss from the butanol production pathway. This was to prevent selection bias for trialing its validity. It is strongly recommended to use the method more flexibly, looking at the FVAp graphs that are produced for the reactions, knowledge of the organism, and the scorings in order to decide on suitable knockouts.

TABLE 5 | Molar ratio of glucose:butanol produced in predicted strains.

Method	Molar ratio (glucose:butanol)
OptKnock	1:0
RobustKnock	1:0
RobOKoD	1:0.9
Experimental	1:0.8

Discussion

These results illustrate two limitations of OptKnock and RobustKnock. First, the knockout predictions are deterministic, not ranked, and a unique set of knockouts is predicted. As shown by these results, different knockouts which may give similar phenotypes cannot be identified by these algorithms. With RobOKoD, a score is attributed to each modification, and one can readily check whether some modifications are expected to result in similar phenotypes and select those that can be more easily implemented experimentally. Secondly, OptKnock and RobustKnock are unable to predict over-expression or dampening strategies, which are of prime interest to increasing or decreasing flux down key pathways, respectively. However, it is argued that using a range of available techniques may help to build up a more comprehensive understanding of the system, and comparing the results obtained by different methods (e.g., Burgard et al., 2003; Choi et al., 2010; Tepper and Shlomi, 2010; Park et al., 2012) would be the most valuable strategy for designing producing strains.

It is also important to note that constraint-based modeling is not appropriate in all instances for prediction of suitable strains for target molecule production. FBA, a key method of assessing the functionality of a given strain, has the flaw whereby side reactions are not predicted to be carrying flux *in silico* as this would reduce the optimal resources that are routed to growth. An example being FBA run on yeast not producing ethanol under an intuitively appealing set of constraints (Westerhoff et al., 2009). This means that only solutions for target production pathways which are heavily coupled with growth can be identified. This is not an issue in most cases since a viable strain is desired but limits the applicability of this framework in particular cases, for example, when there is a need to decouple production from growth. It also means that the false negative rate for *in silico* strain predictions is high, with many successful laboratory strains not appearing so when translated to an *in silico* model. In future the field needs to look more toward different ways of predicting metabolic fluxes. Combining kinetic and stoichiometric models of the metabolic system (Chowdry et al., 2014) provides additional levels of constraints (including enzyme inhibition and activation) and is expected to improve the prediction of effective interventions. A longer term goal is therefore the production of detailed, large-scale kinetic models of the whole metabolic system (Stanford et al., 2013).

When running OptKnock and RobustKnock, it was clear that OptKnock was more user friendly, owing to it being made available in the COBRA Toolbox for Matlab and therefore applicable to a number of MILP (mixed integer linear programming) solvers. This was not the case for RobustKnock, which required a non-standardized model structure and the use of a specific solver, Tomlab, which has limited free access. An additional goal of designing RobOKoD was therefore to ensure its accessibility and robustness by reusing freely-accessible solvers, extensively validated COBRA Toolbox methods, and standardized model formats such as SBML.

A necessary future direction for both RobOKoD and existing tools such as OptKnock and RobustKnock will be to move to making predictions regarding knockouts, over-expressions, etc. at the level of the *gene*, rather than, as currently, at the level of the reaction. Due to the presence of both isoenzymes and promiscuous enzymes, it is clear that there is not a 1:1 mapping between gene and reaction. Consequently, manipulation of a given gene is likely to affect a number of reactions. Modification of this method to consider the gene-protein-reaction (GPR) relationships that are present in many genome-scale metabolic models will be a priority for future development.

To summarize, RobOKoD provides an additional tool to aid the task of designing strains for the (over)production of target products. It is able to predict and rank knockouts, over-expressions, and dampening targets. While predicting an optimized set of gene modifications to implement, unlike other methods, RobOKoD also provides lists of candidate modifications, along with graphical flux variability profiles, allowing the user to manually validate the set of predictions. Such a flexible approach—particularly when used in conjunction with other analysis methods mentioned previously—will allow for sensible gene manipulation approaches to be taken into the laboratory.

Author Contributions

NJS conceived the study, led the project, developed the method and the code, and wrote and led the writing of the manuscript. PM contributed to the method conception and development, and writing of the manuscript. NS contributed to the code development and the writing of the manuscript.

Acknowledgments

NJS is grateful for funding under grant code BB/M013189/1, and the European Union under the Preparatory Phase Projects in the framework of FP7 (project reference 312455). PM received the support of the INRA (program CJS) and of the European Union in the framework of the Marie-Curie FP7 COFUND People Program, through the award of an AgreenSkills fellowship (under grant agreement n° 267196). NS is grateful to the BBSRC for funding under grant code BB/K019783/1. All authors thank both Ettore Murabito and Michael Howard for useful discussions whilst developing this work.

Supplementary Material

The Supplementary Material for this article can be found online at: http://www.frontiersin.org/journal/10.3389/fcell.2015.00017/abstract

References

Angermayr, S. A., van der Woude, A. D., Correddu, D., Vreugdenhil, A., Verrone, V., and Hellingwerf, K. J. (2014). Exploring metabolic engineering design principles for the photosynthetic production of lactic acid by *Synechocystis* sp. PCC6803. *Biotechnol. Biofuels* 7:99. doi: 10.1186/1754-6834-7-99

Atsumi, S., Higashide, W., and Liao, J. C. (2009). Direct photosynthetic recycling of carbon dioxide to isobutyraldehyde. *Nat. Biotechnol.* 27, 1177–1180. doi: 10.1038/nbt.1586

Ballerstein, K., von Kamp, A., Klamt, S., and Haus, U. U. (2012). Minimal cut sets in a metabolic network are elementary modes in a dual network. *Bioinformatics* 28, 381–387. doi: 10.1093/bioinformatics/btr674

Büchel, F., Rodriguez, N., Swainston, N., Wrzodek, C., Czauderna, T., Keller, R., 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

Burgard, A. P., Pharkya, P., and Maranas, C. D. (2003). Optknock: a bilevel programming framework for identifying gene knockout strategies for microbial strain optimization. *Biotechnol. Bioeng.* 84, 647–657. doi: 10.1002/bit.10803

Choi, H. S., Lee, S. Y., Kim, T. Y., and Woo, H. M. (2010). *In silico* identification of gene amplification targets for improvement of lycopene production. *Appl. Environ. Microbiol.* 76, 3097–3105. doi: 10.1128/AEM.00115-10

Chowdry, A., Zomorrodi, A. R., and Maranas, C. D. (2014). k-OptForce: integrating kinetics with flux balance analysis for strain design. *PLoS Comput. Biol.* 10:e1003487. doi: 10.1371/journal.pcbi.1003487

de Figueiredo, L. F., Podhorski, A., Rubio, A., Kaleta, C., Beasley, J. E., Schuster, S., et al. (2009). Computing the shortest elementary flux modes in genome-scale metabolic networks. *Bioinformatics* 25, 3158–3165. doi: 10.1093/bioinformatics/btp564

Dellomonaco, C., Clomburg, J. M., Miller, E. N., and Gonzalez, R. (2011). Engineered reversal of the β-oxidation cycle for the synthesis of fuels and chemicals. *Nature* 476, 355–359. doi: 10.1038/nature10333

Dobson, P. D., Smallbone, K., Jameson, D., Simeonidis, E., Lanthaler, K., Pir, P., et al. (2010). Further developments towards a genome-scale metabolic model of yeast. *BMC Syst. Biol.* 4:145. doi: 10.1186/1752-0509-4-145

Feist, A. M., Henry, C. S., Reed, J. L., Krummenacker, M., Joyce, A. R., Karp, P. D., et al. (2007). A genome-scale metabolic reconstruction for *Escherichia coli* K-12 MG1655 that accounts for 1260 ORFs and thermodynamic information. *Mol. Syst. Biol.* 3:121. doi: 10.1038/msb4100155

Feist, A. M., Zielinski, D. C., Orth, J. D., Schellenberger, J., Herrgard, M. J., and Palsson, B. Ø. (2010). Model-driven evaluation of the production potential for growth-coupled products of *Escherichia coli*. *Metab. Eng.* 12, 173–186. doi: 10.1016/j.ymben.2009.10.003

Gudmundsson, S., and Thiele, I. (2010). Computationally efficient flux variability analysis. *BMC Bioinformatics* 11:489. doi: 10.1186/1471-2105-11-489

Harcombe, W. R., Delaney, N. F., Leiby, N., Klitgord, N., and Marx, C. J. (2013). The ability of flux balance analysis to predict evolution of central metabolism scales with the initial distance to the optimum. *PLoS Comput. Biol.* 9:e1003091. doi: 10.1371/journal.pcbi.1003091

Heavner, B. D., Smallbone, K., Barker, B., Mendes, P., and Walker, L. P. (2012). Yeast 5 – an expanded reconstruction of the *Saccharomyces cerevisiae* metabolic network. *BMC Syst. Biol.* 6:55. doi: 10.1186/1752-0509-6-55

Heiner, M. (2009). "Understanding network behavior by structured representations of transition invariants," in *Algorithmic Bioprocesses: Natural Computing Series*, eds A. Condon, D. Harel, J. N. Kok, A. Salomaa, and E. Winfree (Berlin; Heidelberg: Springer-Verlag), 367–389.

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

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

Kitano, H. (2002). Systems biology: a brief overview. *Science* 295, 1662–1664. doi: 10.1126/science.1069492

Koide, T., Pang, W. L., and Baliga, N. S. (2009). The role of predictive modelling in rationally re-engineering biological systems. *Nat. Rev. Microbiol.* 7, 297–305. doi: 10.1038/nrmicro2107

Lee, D., Smallbone, K., Dunn, W. B., Murabito, E., Winder, C. L., Kell, D. B., et al. (2012). Improving metabolic flux predictions using absolute gene expression data. *BMC Syst. Biol.* 6:73. doi: 10.1186/1752-0509-6-73

Li, X., Guo, D., Cheng, Y., Zhu, F., Deng, Z., and Liu, T. (2014). Overproduction of fatty acids in engineered *Saccharomyces cerevisiae*. *Biotechnol. Bioeng.* 111, 1841–1852. doi: 10.1002/bit.25239

Liao, Y. C., Huang, T. W., Chen, F. C., Charusanti, P., Hong, J. S., Chang, H. Y., et al. (2011). An experimentally validated genome-scale metabolic reconstruction of *Klebsiella pneumoniae* MGH 78578, iYL1228. *J. Bacteriol.* 193, 1710–1717. doi: 10.1128/JB.01218-10

Lo, T. M., Teo, W. S., Ling, H., Chen, B., Kang, A., and Chang, M. W. (2013). Microbial engineering strategies to improve cell viability for biochemical production. *Biotechnol. Adv.* 31, 903–914. doi: 10.1016/j.biotechadv.2013.02.001

Ng, C. Y., Jung, M.-Y., Lee, J., and Oh, M.-K. (2012). Production of 2,3-butanediol in *Saccharomyces cerevisiae* by *in silico* aided metabolic engineering. *Microb. Cell Fact.* 11:68. doi: 10.1186/1475-2859-11-68

Orth, J. D., Thiele, I., and Palsson, B. Ø. (2010). What is flux balance analysis? *Nat. Biotechnol.* 28, 245–248. doi: 10.1038/nbt.1614

Park, J. M., Park, H. M., Kim, W. J., Kim, H. U., Kim, T. Y., and Lee, S. Y. (2012). Flux variability scanning based on enforced objective flux for identifying gene amplification targets. *BMC Syst. Biol.* 6:106. doi: 10.1186/1752-0509-6-106

Pharkya, P., and Maranas, C. D. (2006). An optimization framework for identifying reaction activation/inhibition or elimination candidates for overproduction in microbial systems. *Metab. Eng.* 8, 1–13. doi: 10.1016/j.ymben.2005.08.003

Rocha, I., Maia, P., Evangelista, P., Vilaça, P., Soares, S., Pinto, J. P., et al. (2010). OptFlux: an open-source software platform for *in silico* metabolic engineering. *BMC Syst. Biol.* 4:45. doi: 10.1186/1752-0509-4-45

Sauer, U. (2006). Metabolic networks in motion: ^{13}C-based flux analysis. *Mol. Syst. Biol.* 2, 62. doi: 10.1038/msb4100109

Schellenberger, J., Que, R., Fleming, R. M., Thiele, I., Orth, J. D., Feist, A. M., et al. (2011). Quantitative prediction of cellular metabolism with constraint-based models: the COBRA Toolbox v2.0. *Nat. Protoc.* 6, 1290–1307. doi: 10.1038/nprot.2011.308

Stanford, N. J., Lubitz, T., Smallbone, K., Klipp, E., Mendes, P., and Liebermeister, W. (2013). Systematic construction of kinetic models from genome-scale metabolic networks. *PLoS ONE* 8:e79195. doi: 10.1371/journal.pone.0079195

Tepper, N., and Shlomi, T. (2010). Predicting metabolic engineering knockout strategies for chemical production: accounting for competing pathways. *Bioinformatics* 26, 536–543. doi: 10.1093/bioinformatics/btp704

Thiele, I., and Palsson, B. Ø. (2010). A protocol for generating a high-quality genome-scale metabolic reconstruction. *Nat. Protoc.* 5, 93–121. doi: 10.1038/nprot.2009.203

Varman, A. M., Xiao, Y., Leonard, E., and Tang, Y. J. (2011). Statistics-based model for prediction of chemical biosynthesis yield from *Saccharomyces cerevisiae*. *Microb. Cell Fact.* 10:45. doi: 10.1186/1475-2859-10-45

von Kamp, A., and Klamt, S. (2014). Enumeration of smallest intervention strategies in genome-scale metabolic networks. *PLoS Comput. Biol.* 10:e1003378. doi: 10.1371/journal.pcbi.1003378

Westerhoff, H. V., Winder, C., Messiha, H., Simeonidis, E., Adamczyk, M., Verma, M., et al. (2009). Systems biology: the elements and principles of life. *FEBS Lett.* 583, 3882–3890. doi: 10.1016/j.febslet.2009.11.018

Yuan, Y., Bi, C., Nicolaou, S. A., Zingaro, K. A., Ralston, M., and Papoutsakis, E. T. (2014). Overexpression of the *Lactobacillus plantarum* peptidoglycan biosynthesis *murA2* gene increases the tolerance of *Escherichia coli* to alcohols and enhances ethanol production. *Appl. Microbiol. Biotechnol.* 98, 8399–8411. doi: 10.1007/s00253-014-6004-0

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.

Reaction Abbreviations

Model reaction ID	Reaction name	EC
ACALD	Acetaldehyde dehydrogenase (acetylating)	1.2.1.10
er_027	Alcohol dehydrogenase (to butanol)	1.1.1.1
LDH_D	D-lactate dehydrogenase	1.1.1.27
NADH16	NADH dehydrogenase (ubiquinone)	1.6.5.3
NH4t	Ammonia reversible transport	n/a
O2t	O_2 transport via diffusion	n/a
PFL	Pyruvate formate lyase	2.3.1.54
PIt2r	Phosphate reversible transport via proton symport	n/a

frontiers in
BIOENGINEERING AND BIOTECHNOLOGY

ORIGINAL RESEARCH ARTICLE
published: 05 January 2015
doi: 10.3389/fbioe.2014.00076

Succinate overproduction: a case study of computational strain design using a comprehensive *Escherichia coli* kinetic model

Ali Khodayari*†, *Anupam Chowdhury*† and *Costas D. Maranas*

Department of Chemical Engineering, The Pennsylvania State University, University Park, PA, USA

Edited by:
Daniel Machado, University of Minho, Portugal

Reviewed by:
Cong T. Trinh, The University of Tennessee Knoxville, USA
Joonhoon Kim, University of Wisconsin-Madison, USA
Kiran Raosaheb Patil, European Molecular Biology Laboratory, Germany

***Correspondence:**
Costas D. Maranas, Department of Chemical Engineering, The Pennsylvania State University, 112 Fenske Laboratory, University Park, PA 16802, USA
e-mail: costas@psu.edu

†Joint first authors

Computational strain-design prediction accuracy has been the focus for many recent efforts through the selective integration of kinetic information into metabolic models. In general, kinetic model prediction quality is determined by the range and scope of genetic and/or environmental perturbations used during parameterization. In this effort, we apply the k-OptForce procedure on a kinetic model of *E. coli* core metabolism constructed using the Ensemble Modeling (EM) method and parameterized using multiple mutant strains data under aerobic respiration with glucose as the carbon source. Minimal interventions are identified that improve succinate yield under both aerobic and anaerobic conditions to test the fidelity of model predictions under both genetic and environmental perturbations. Under aerobic condition, k-OptForce identifies interventions that match existing experimental strategies while pointing at a number of unexplored flux re-directions such as routing glyoxylate flux through the glycerate metabolism to improve succinate yield. Many of the identified interventions rely on the kinetic descriptions that would not be discoverable by a purely stoichiometric description. In contrast, under fermentative (anaerobic) condition, k-OptForce fails to identify key interventions including up-regulation of anaplerotic reactions and elimination of competitive fermentative products. This is due to the fact that the pathways activated under anaerobic condition were not properly parameterized as only aerobic flux data were used in the model construction. This study shed light on the importance of condition-specific model parameterization and provides insight on how to augment kinetic models so as to correctly respond to multiple environmental perturbations.

Keywords: computational strain design, kinetic model, bilevel optimization, succinate overproduction, model parameterization

INTRODUCTION

Engineered microorganisms are increasingly being used as cellular factories for the bio-production of chemicals of interest (Curran and Alper, 2012; Hong and Nielsen, 2012; Lee et al., 2012). Keeping pace with genome editing techniques for strain design, several computational tools have been developed to identify system-wide genetic modification strategies that improve the yield of targeted biochemicals (Pharkya et al., 2004; Kim et al., 2011; Xu et al., 2011; Maia et al., 2012; Cotten and Reed, 2013a). In general, these tools rely on a stoichiometric representation of a metabolic network and solve bilevel optimization problems to suggest prioritized intervention strategies that divert metabolic flux towards the chemical of interest (Segre et al., 2002; Burgard et al., 2003; Kim and Reed, 2010; Rocha et al., 2010; Tepper and Shlomi, 2010). The methodology and comparative benefits of each procedure is discussed in detail elsewhere (Zomorrodi et al., 2012). However, key methodological impediments of these approaches are the stoichiometry-only representation of metabolism and the on–off representation of regulation. This may lead to a metabolite concentration, enzymatic activity, and metabolic regulation-agnostic intervention strategies. Therefore, identified flux re-direction predictions (especially up/down flux modulation) are sometimes difficult to translate into actionable genetic interventions. For example, it is unclear if a desired metabolic flux up-regulation is achievable or even consistent with enzyme kinetics or physiological metabolite concentrations.

Some of the shortcomings of genome-scale stoichiometric models in quantifying the effect of concentration and enzyme levels on reaction throughput and regulation can be addressed by kinetic models of metabolism (Mahadevan et al., 2002; Fleming et al., 2010; Jamshidi and Palsson, 2010; Smallbone et al., 2010; Feng et al., 2012). Kinetic models yield a system of ordinary differential equations (ODEs) that describe the time evolution of metabolite concentrations, enzyme activities, and reaction fluxes. Several efforts have been made in recent years for improving the accuracy of stoichiometry-based tools by partially integrating kinetic information (Nikolaev, 2010; Song and Ramkrishna, 2012; Angermayr and Hellingwerf, 2013; Almquist et al., 2014). However, most of these procedures are aimed towards improved metabolic phenotype prediction through *ad hoc* constraints (Cotten and Reed, 2013b) rather than strain design. The k-OptForce procedure (Chowdhury et al., 2014) extends the previously developed strain-design OptForce algorithm (Ranganathan et al., 2010) by integrating all available mechanistic details afforded by kinetic models within a constraint-based optimization framework tractable even for genome-scale models. Reactions with available kinetic descriptions yield (generally unique) steady-state flux values while the remaining reactions are only constrained by stoichiometric relations. Genetic intervention strategies consistent with restrictions imposed by maximum enzyme activity, bounds on metabolite concentrations and kinetic expressions are identified using a bilevel Mixed Integer Nonlinear Program (MINLP) optimization framework (Chowdhury et al., 2014). Examples addressed in Chowdhury et al. (2014), however, accounted for only a handful of reactions with kinetic expressions.

In this paper, we apply k-OptForce procedure for the recently published large-scale kinetic model of *E. coli* core metabolism (Khodayari et al., 2014). The kinetic model includes 138 reactions, 93 metabolites, and 60 substrate-level regulatory interactions and accounts for glycolysis/gluconeogenesis, pentose phosphate (PP) pathway, TCA cycle, major pyruvate metabolism, anaplerotic reactions, glyoxylate shunt, Entner–Doudoroff (ED) pathway, and a number of reactions in other parts of the metabolism. The model was parameterized using the ensemble modeling (EM) formalism (Tran et al., 2008) by simultaneously satisfying normalized flux data per 100 mmol of glucose uptake (for approximately 25 reactions per mutant) for the wild-type and seven single gene deletion mutants, under aerobic condition (Ishii et al., 2007). The EM approach decomposes all reactions into elementary steps bypassing the need of detail kinetic expressions. First, an ensemble of kinetic models is generated by uniformly sampling reaction reversibilities and enzyme fractions following different time trajectories but all reaching the same steady-state flux values (Tan and Liao, 2012). Next, a Genetic Algorithm (GA) implementation is used to "swap" kinetic parameterizations between models in the ensemble so as to minimize the deviations from all set of mutant network fluxes. Models constructed using flux data for a single strain do not always perform well in predicting deletion strain metabolic phenotypes (Jouhten, 2012; Villaverde et al., 2014). Unlike stoichiometric models that could reveal physiologically relevant flux re-directions in response to perturbations by re-optimizing biomass yield, kinetic models must be endowed beforehand with all known substrate-level regulatory interactions to capture metabolic responses to genetic/environmental perturbations (Jouhten, 2012; Heijnen and Verheijen, 2013; Villaverde et al., 2014). Note that while the EM based elementary mode analysis was used for strain design in an earlier effort (Flowers et al., 2013), the limited scope of the model may fail to capture genome-scale flux re-directions.

The k-OptForce procedure (Chowdhury et al., 2014) was used to identify the minimal interventions that maximize the yield of succinate production using a hybrid kinetic (Khodayari et al., 2014) and stoichiometric *i*AF1260 (Feist et al., 2007) description of *E. coli* metabolism. Succinate was chosen as the target bioproduct as there exists numerous experimental strain-engineering studies to compare the suggestions of k-OptForce procedure (Lee et al., 2005; Cao et al., 2011; Tan et al., 2011). This study was carried out under both aerobic and anaerobic conditions to assess the fidelity of the kinetic model when used to make predictions for a different environmental condition (i.e., anaerobic) than the one parameterized for (i.e., aerobic). The goal was to quantify the reduction in prediction quality moving from aerobic to anaerobic under glucose minimal condition and suggest model modifications that remedy these shortcomings. k-OptForce recapitulated existing strategies while also pointing at promising but currently unexplored interventions. In addition, results under anaerobic condition indicate that the kinetic model needs to be re-parameterized with mutant flux information involving a reversed TCA cycle routing flux towards succinate. A number of regulatory modifications of the kinetic model are also found to be necessary to better reflect metabolic fluxes associated with anaerobic succinate production. These include activation of fermentation pathways and pyruvate formate lyase (PFL) by key regulatory proteins FNR (fumarate and nitrate reductase regulation) and ArcA (aerobic respiratory control).

MATERIALS AND METHODS

Using k-OptForce, the genome-scale stoichiometry matrix is divided into two parts: reactions with stoichiometric information only ($\boldsymbol{J}^{\text{stoic}}$), and those having additional kinetic information ($\boldsymbol{J}^{\text{kin}}$). A schematic representation of the framework is depicted in **Figure 1**. The kinetic information was extracted from the kinetic model of *E. coli* central metabolism developed in Khodayari et al. (2014). The number of reactions in the kinetic representation is a compromise between reduction of solution space using kinetic data and run time for solving the non-linear expressions of mass conservations. Upon exclusion of the exchange/transport reactions and elimination of reactions not involved in succinate synthesis (such as glycogen pathway), a subset of the kinetic model was selected containing 36 reactions and 31 metabolites. The resulting model includes reactions from glycolysis/gluconeogenesis, PP pathway, TCA cycle, anaplerotic reactions, glyoxylate shunt, and ED pathway with available experimental data during model parameterization. This model was finally supplemented with the stoichiometric *i*AF1260 model of *E. coli* (Feist et al., 2007).

Glucose minimal condition were simulated by restricting glucose uptake flux (which serves as a basis for the fluxes in the metabolic network) to −100 mmol $\text{gDW}^{-1}\text{h}^{-1}$. Oxygen uptake was limited to −200 mmol $\text{gDW}^{-1}\text{h}^{-1}$ for aerobic condition and set to zero for fermentative condition. Regulatory information for both aerobic and anaerobic conditions was imported from the supplementary material of *i*AF1260 model (Feist et al., 2007). The minimum production levels of succinate was set at 90% of its theoretical maximum for each condition (i.e., 135 mmol $\text{gDW}^{-1}\text{h}^{-1}$ in aerobic and 149 mmol $\text{gDW}^{-1}\text{h}^{-1}$ in anaerobic conditions) while a minimum level of biomass production equal to 10% of its theoretical maximum was simultaneously imposed (i.e., 0.965 h^{-1} in aerobic and 0.303 h^{-1} in anaerobic conditions). The k-OptForce algorithm was implemented in the same stepwise procedure as described previously [see Methods in Chowdhury et al. (2014) for details]. At first, we identify all reactions that must depart (hence called MUST sets) from the reference phenotype to realize the targeted levels of overproduction of the desired chemicals under stoichiometric and kinetic constraints. Subsequently, we solve a bilevel optimization formulation (see **Figure 1E**) where we maximize the target flux by gradually increasing the total number (κ) of enzymatic interventions (for reactions in $\boldsymbol{J}^{\text{kin}}$) and/or flux manipulations (for reactions in $\boldsymbol{J}^{\text{stoic}}$) from the MUST sets. Starting from

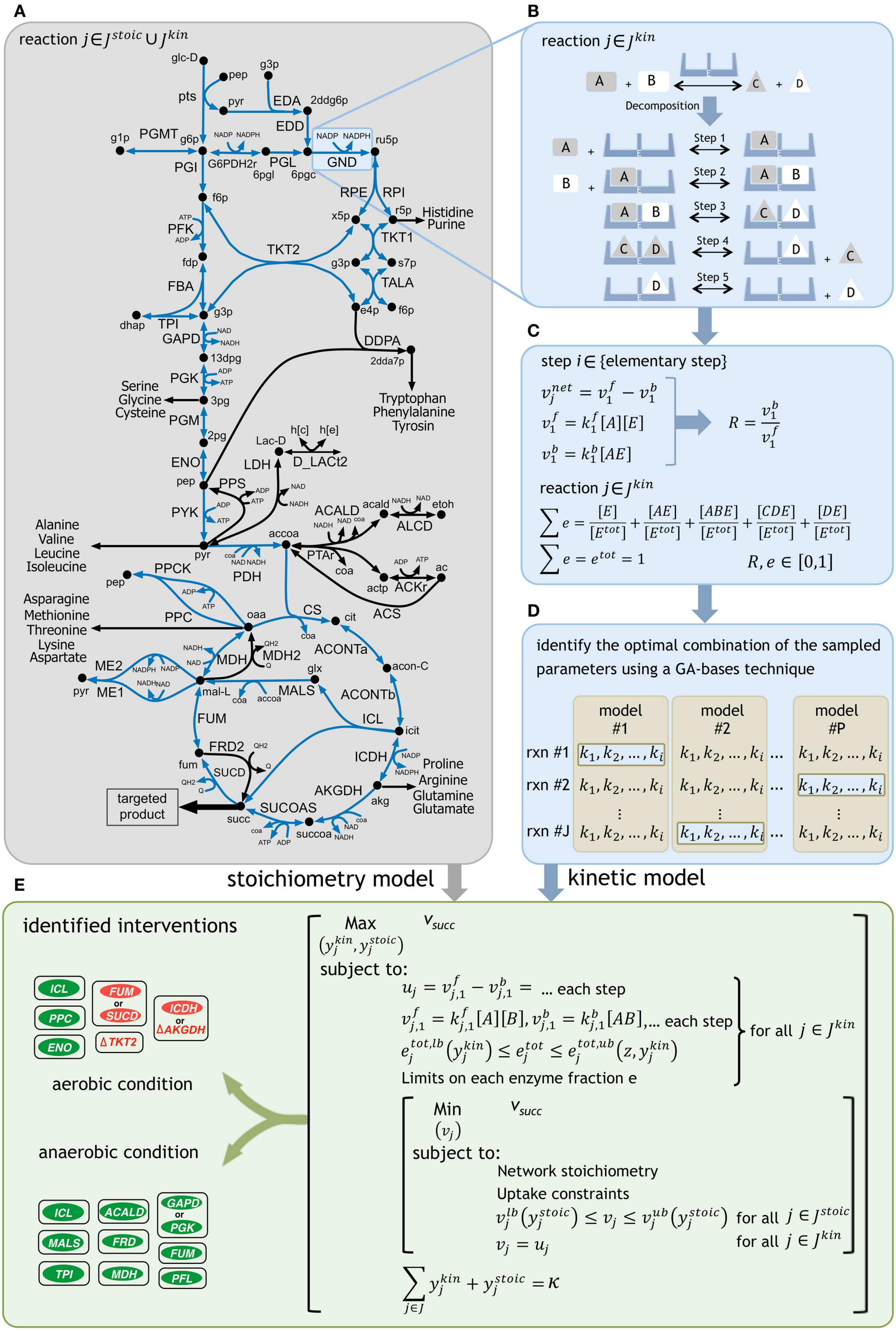


FIGURE 1 | A schematic representation of the framework.

(Continued)

FIGURE 1 | Continued

(A) The reactions with kinetic descriptions are shown in blue. **(B)** The reactions are first decomposed into their elementary steps. **(C)** Elementary kinetic parameters are expressed as a function of reaction reversibilities and enzyme fractions. Reaction reversibilities and enzyme fractions are sampled to construct an ensemble of models, for any given reaction. **(D)** A genetic algorithm (GA) implementation identifies the optimal combination of the sampled parameters by minimizing the deviation from experimentally measured flux data for multiple mutant strains [see Methods of Khodayari et al. (2014)]. **(E)** The k-OptForce procedure identifies a minimal set of interventions that maximizes the yield of targeted product [see Methods of Chowdhury et al. (2014)].

a single intervention, we stop this procedure when the target flux does not improve appreciably with additional interventions. The optimization formulations for the characterization of the overproducing network and identification of the FORCE sets were altered from the original procedure to incorporate the kinetic information of each reaction in $\boldsymbol{J}^{\text{kin}}$ as a function of the decomposed expressions of its elementary steps (see **Figure 1**) instead of directly manipulating the reaction enzyme activities ($v^{\max}$). Additional constraints were imposed to express the flux of each reaction in $\boldsymbol{J}^{\text{kin}}$ as the difference of the forward and reverse reactions for each elementary step. The sum of individual enzyme fractions e is represented by e^{tot} (i.e., normalized total enzyme concentration) that is equal to one in the reference (wild-type) strain, but varies when up/down-regulated in mutant strains. Here, we allowed the e^{tot} of each reaction to vary between zero (i.e., removal of its activity) and a 10-fold up-regulation in its expression to account for individual enzymatic perturbations in mutant strains. Likewise, the same limits of variation were set for the individual enzyme fractions e for each reaction.

The metabolite concentrations were allowed to vary within fivefold from their steady-state values in the reference phenotype. The FORCE set of interventions was identified in a two-step procedure [see Methods of Chowdhury et al. (2014)]. The first step identified the reactions in $\boldsymbol{J}^{\text{kin}}$ (using binary variables y^{kin}) whose enzymatic activity (i.e., e^{tot}) must be altered from their reference level to achieve the required flux re-distribution towards succinate. The lower and upper bounds on e^{tot} (i.e., $e^{\text{tot,lb}}$ and $e^{\text{tot,ub}}$) are functions of y^{kin} and the maximum fold-change z, as follows:

$$e_j^{\text{tot,lb}} = \begin{cases} 1, & \text{if } j \in J^{\text{kin}} \backslash \text{MUST}^L \\ 1 - y_j^{\text{kin}}, & \text{if } j \in J^{\text{kin}} \cap \text{MUST}^L \end{cases}$$

$$e_j^{\text{tot,ub}} = \begin{cases} 1, & \text{if } j \in J^{\text{kin}} \backslash \text{MUST}^U \\ (z-1)\, y_j^{\text{kin}} + 1, & \text{if } j \in J^{\text{kin}} \cap \text{MUST}^U \end{cases}$$

If selected for down-regulation (i.e., when the reaction is part of MUST^L), e^{tot} is allowed to vary from zero ($e^{\text{tot,lb}} = 0$ for $y^{\text{kin}} = 1$) to its reference expression. Otherwise, e^{tot} is set to one. Likewise, if selected for up-regulation (i.e., when the reaction is part of MUST^U), e^{tot} is allowed to vary from one to a z-fold overexpression ($e^{\text{tot,ub}} = z$ for $y^{\text{kin}} = 1$). The MINLP formulation for the first-step was initially solved using a local solver [DICOPT (Grossmann et al., 2002)], and the results were used as inputs to find the global optimum using the BARON solver (Sahinidis, 1996). Subsequently, by fixing the fluxes in $\boldsymbol{J}^{\text{kin}}$, the second step identified additional flux manipulations in $\boldsymbol{J}^{\text{stoic}}$ (using binary variables y^{stoic}) while assuming a worst-case scenario for the inner objective function. The relation of the modified bounds $\left(v_j^{\text{lb}}, v_j^{\text{ub}}\right)$ on the reaction fluxes in $\boldsymbol{J}^{\text{stoic}}$ with y^{stoic} is similar to that explained for the first step of FORCE set identification for the implementation of up/down-regulations and/or reaction removals [see Methods of Chowdhury et al. (2014)].

RESULTS

EXAMINING THE PREDICTIVE PERFORMANCE OF THE KINETIC MODEL

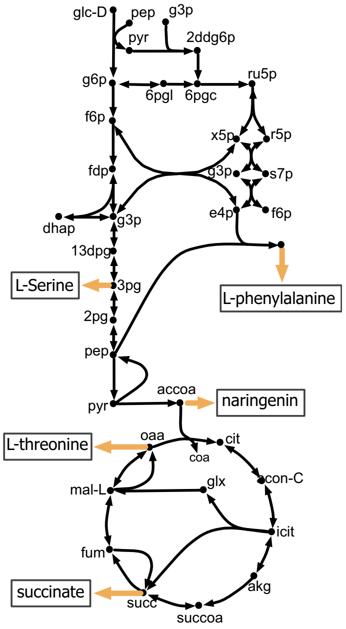
The perturbed phenotype prediction accuracy of the parameterized kinetic model was first assessed for five different engineered strains under aerobic condition. The experimentally reported product yield was compared against the kinetic model and FBA predictions (see **Table 1**). A twofold up-regulation for small fold-change, and 10-fold up-regulation for a high fold-change are used to express enzyme up-regulation, whenever such information is not available in the relevant literature. The enzyme level manipulation in the kinetic model is achieved by changing e^{tot} for each particular enzyme. Gene deletions are implemented by setting the e^{tot} of the encoded enzyme to zero.

The kinetic model closely matches the succinate producing strain while FBA over-estimates it because the former captures the feed-forward inhibition on glyoxylate shunt by intermediates phosphoenolpyruvate (pep) (MacKintosh and Nimmo, 1988; Ogawa et al., 2007) and isocitrate (icit) (Hoyt et al., 1988). For both L-serine and L-threonine, FBA directs all carbon flux towards biomass predicting little to no amount of product formation. The kinetic model over-estimates L-serine yield as product inhibition of the phosphoglycerate dehydrogenase (PGCD) (Grant, 2012; Li et al., 2012; Wang et al., 2014) is not captured in the kinetic model (see **Figure 2A**). In contrast, the kinetic model under-estimates the yield of L-phenylalanine production. A possible reason is that the feed-forward activation of pep on 5-enolpyruvylshikimate-3-phospahte synthase (EPSPS) (Gruys et al., 1992) is absent in the kinetic model (see **Figure 2B**). In addition, due to lack of experimental data during parameterization, the model over-estimates the inhibitory effect of phosphate on transaldolase (TALA) activity (Sprenger et al., 1995), which further restricts flux towards L-phenylalanine production. The naringenin engineered strain productivity is better reflected by the kinetic model as FBA does not capture the feedback inhibition of acetyl-CoA on phosphoglucomutase (PGM) activity (Sanwal et al., 1972; Duckworth et al., 1973) that limits flux towards the flavanone pathway.

OVERPRODUCTION OF SUCCINATE UNDER AEROBIC CONDITION

Both OptForce and k-OptForce adopt similar strategies for re-directing flux towards succinate under aerobic condition by routing more flux through isocitrate lyase (ICL), increasing flux through phosphoenolpyruvate carboxylase (PPC), and converting the intermediate glyoxylate back to glycerate 2-phosphate (2pg) using glycerate metabolism (see **Figure 3**). However, the number of required interventions varies. While OptForce suggests that only three interventions are required to achieve a

Table 1 | A comparison between model predictions and experimental yields for five different products in *E. coli* under aerobic condition.



Target product	Interventions with enzyme-fold-change	Yield (mol product/mol glucose)		
		FBA	Kinetic model	Experimental data
Succinate	ΔSUCD ICL 10↑ PPC 2↑	0.99	0.52	0.6 (Lin et al., 2005b)
L-serine	ΔPDH PGCD 10↑ PGK 2↑	0–0.01	0.81	0.48 (Lai et al., 2012)
L-threonine	PPC 2↑ ICL 2↑	0–0.04	0.52	0.59 (Lee et al., 2007)
L-phenyl alanine	ΔPYK DDPA 10↑ TKT1 10↑	0.44	0.11	0.36 (Baez-Viveros et al., 2007)
Naringenin	ΔSUCOAS ΔFUM ACCOAC 10↑ PDH 10↑ GAPD 10↑	0.43	0.07	0.11 (Xu et al., 2011)

The engineering strains are simulated using both the kinetic model and FBA (max biomass).

SUCD, succinate dehydrogenase; ICL, isocitrate lyase; PPC, phosphoenolpyruvate carboxylase; PDH, pyruvate dehydrogenase; PGCD, phosphoglycerate dehydrogenase; PGK, phosphoglycerate kinase; PPC, phosphoenolpyruvate carboxylase; PYK, pyruvate kinase; DDPA, 3-deoxy-D-arabino-heptulosonate 7-phosphate synthetase; TKT, transketolase; SUCOAS, succinyl-CoA synthetase; FUM, fumarase; ACCOAC, acetyl-CoA carboxylase; GAPD, glyceraldehyde-3-phosphate dehydrogenase.

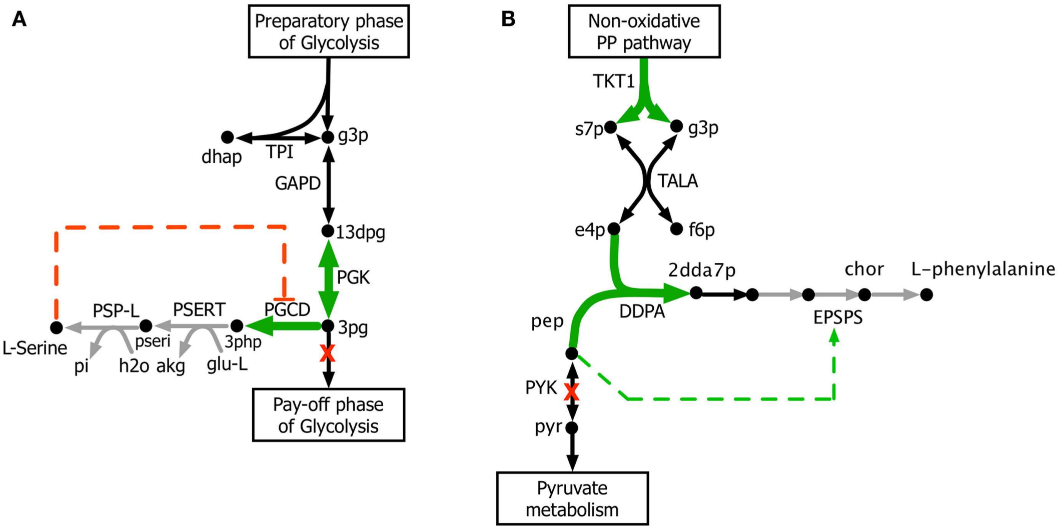


FIGURE 2 | Biosynthesis pathways for (A) L-serine and (B) L-phenylalanine. The suggested up-regulations and knock-outs are shown with green color and red crosses, respectively. The reactions absent in the current kinetic model are shown in gray. Missing regulatory interactions (i.e., activation and inhibition) are shown with dashed lines.

succinate yield of 90% of its theoretical maximum, k-OptForce suggests that additional direct up-regulations in the glycolysis and TCA cycle are necessary. For example, k-OptForce suggests at least ninefold up-regulation of ICL enzyme activity to pull TCA cycle flux from icit towards succinate. Likewise, up-regulation of enolase (ENO) enzyme by twofold of its reference activity is required to push more glycolytic flux towards succinate precursors oxaloacetate (oaa) and acetyl-CoA. Regular OptForce suggests that up-regulation of aconitase (ACONT) and down-regulation of isocitrate dehydrogenase (ICDH) are sufficient to indirectly increase flux through PPC and ICL. In contrast, k-OptForce suggests that PPC and ICL must be directly up-regulated to improve

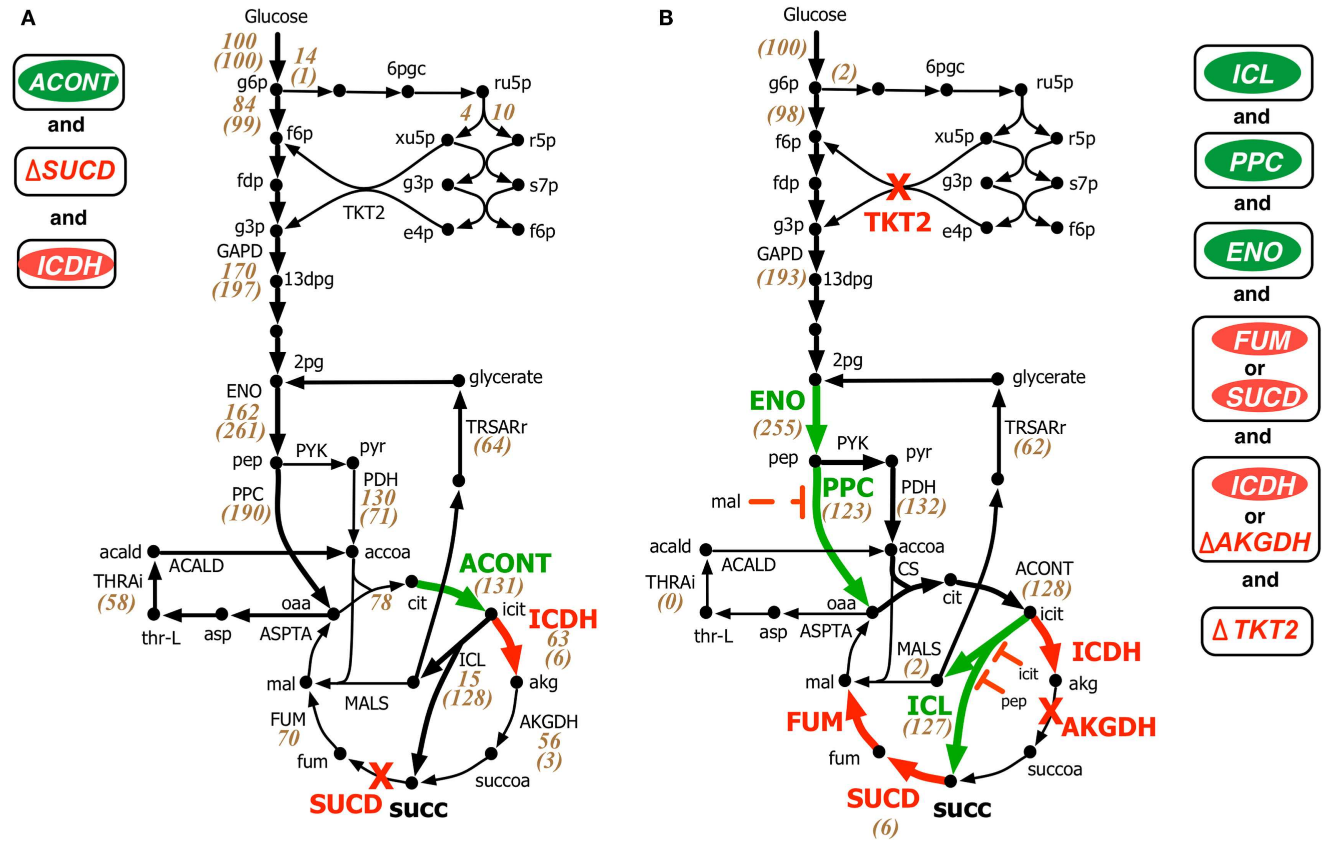


FIGURE 3 | Comparison of intervention strategies predicted by (A) regular OptForce and (B) k-OptForce for overproduction of succinate under aerobic condition in *E. coli*. The values within parentheses indicate the metabolic flux in mmol gDW^{-1} h^{-1} per 100 mmol gDW^{-1} h^{-1} glucose uptake. The values without parentheses in **(A)** show steady-state flux distribution of the reference (wild-type) strain used for kinetic model parameterization (Ishii et al., 2007).

succinate yield. In addition, up-regulation of ENO pulls glyoxylate flux towards 2pg through the glycerate pathway to compensate for the pep depletion. OptForce does not require any enzymatic intervention to route metabolic flux towards acetyl-CoA sending a significant portion (58 mmol gDW^{-1}h^{-1}) from oaa towards acetyl-CoA using the threonine pathway. k-OptForce reveals that such a high flux cannot be routed through the threonine pathway. Even with maximum (i.e., 10-fold) up-regulation of the aspartate transaminase (ASPTA) only 20 mmol gDW^{-1}h^{-1} can be diverted towards threonine. In addition, k-OptForce suggests up-regulation of PPC enzyme activity (by 50% of its reference activity) to ensure availability of equal amounts of acetyl-CoA and oaa for the production of citrate thus preventing the accumulation of intermediates.

The abovementioned interventions suggested by k-OptForce are geared towards circumventing upper bounds on max enzyme activities (i.e., no more than 10-fold). However, limits on metabolite concentrations also play a significant role in restricting flux towards succinate. The maximum yield of succinate suggested by k-OptForce (1.2 mol/mol glucose, 80% of theoretical maximum) is less than the one suggested by OptForce (1.3 mol/mol glucose, 90% of theoretical maximum). This is because as ICL is up-regulated, the concentration of intermediates pep and icit increase reaching twice their reference values. As these metabolites are competitive inhibitors of ICL, the maximum flux through the pathway towards succinate is restricted. In addition, to alleviate the regulatory effect of malate (mal) on the activity of PPC, k-OptForce also proposed a 10-fold down-regulation of the enzymes that catalyze mal production, fumarase (FUM), or succinate dehydrogenase (SUCD). Likewise, k-OptForce suggests removal of transketolase (TKT2) to alleviate the inhibition of 6-phospho-D-gluconate (6pgc) on glucose-6-phosphate isomerase (PGI) to improve the glycolytic flux towards succinate, which also reduces the production of biomass precursors.

Most of the k-OptForce interventions were consistent with engineering efforts aimed at improving succinate production under aerobic condition. For example, up-regulation of ICL and removal of SUCD and ICDH activities improved succinate yield in *E. coli* to 0.5 mol/mol glucose (Lin et al., 2005b). Further improvements in succinate production (up to 0.7 mol/mol glucose) have been achieved by up-regulation of PPC (Lin et al., 2005a). Notably, the same interventions improved aerobic succinate production in *C. glutamicum* to 0.5 mol/mol glucose (Litsanov et al., 2012). Similar to proportional up-regulation of ENO and PPC that fixes the branching ratio of the metabolic flux at pep, regulation of pep to pyruvate in the phosphotransferase system (PTS) reaction for glucose uptake was suggested to reduce the accumulation of intermediates (pyruvate and acetate) and improve succinate yield (Lin et al., 2005a). k-OptForce, however, fails to capture the accumulation of acetate upon up-regulation of PPC and glyoxylate shunt (Lin et al., 2005a; Zhu et al., 2013). This may be due to the fact that no fluxomic data for mutant strains with anaplerotic/glyoxylate shunt up-regulations was included during kinetic model parameterization. As a result, the kinetic model is unaware of the up-regulation that leads towards increased acetate production. Interestingly, k-OptForce routes glyoxylate (formed by the ICL reaction) back to 2pg using the glycerate pathway instead of the malate synthase (MALS) reaction. This pathway improves the yield of succinate since it reduces the overall loss of carbon flux to carbon dioxide. This pathway was engineered by *E. coli* (Hubbard et al., 1998; Osterhout et al., 2011) for the production of ethylene glycol and glucarate consumption, respectively, but remains to be explored for succinate overproduction.

OVERPRODUCTION OF SUCCINATE UNDER ANAEROBIC CONDITION

Under fermentative condition the electron transport chain is not active, thus preventing the oxidation of cofactor NADH generated primarily in glyceraldehyde 3-phosphate dehydrogenase (GAPD) reaction in glycolysis back to NAD. Without an adequate NADH sink, significant amount of metabolic flux is routed towards fermentative products such as ethanol, acetate, lactate, formate, etc. to restore redox balance and cellular growth. Therefore, the general strategy for succinate overproduction is to eliminate all competitive fermentative pathways while pushing more flux towards succinate through the glyoxylate shunt and reversing the reductive branch of TCA cycle (see **Figure 4**). This flux re-direction also regenerates NAD, thus simultaneously coupling succinate production with biomass generation.

In contrast to the aerobic case, k-OptForce suggestions for the anaerobic overproduction of succinate are less accurate compared to OptForce predictions. OptForce requires only five interventions to achieve a succinate yield of 1.42 mol/mol glucose. However, k-OptForce suggests a maximum yield of only 1.08 mol/mol glucose even after nine interventions. While k-OptForce recapitulates some of the interventions identified by OptForce (e.g., threefold up-regulation of the glyoxylate pathway enzymes ICL and MALS), the remaining suggestions deviate from OptForce and proven engineering strategies. The sources of these discrepancies can be traced back to incompatible parameterization of the kinetic model for the anaerobic case. First, due to absence of sufficient flux data in the parameterization procedure, the kinetic model was not tuned to capture reversal of the reductive branch of the TCA cycle necessary for succinate overproduction. k-OptForce suggests up-regulation of all three enzymes of the reductive branch [i.e., malate dehydrogenase (MDH), FUM, and fumarate reductase (FRD)]. However, even after a 6.5-fold up-regulation in MDH activity and 10-fold up-regulation in FUM only 80% of the anaplerotic flux (57 mmol gDW^{-1} h^{-1}) goes towards succinate, while the remaining amount (11 mmol gDW^{-1} h^{-1}) uses the aspartate metabolism to bypasses MDH and FUM (see **Figure 4B**).

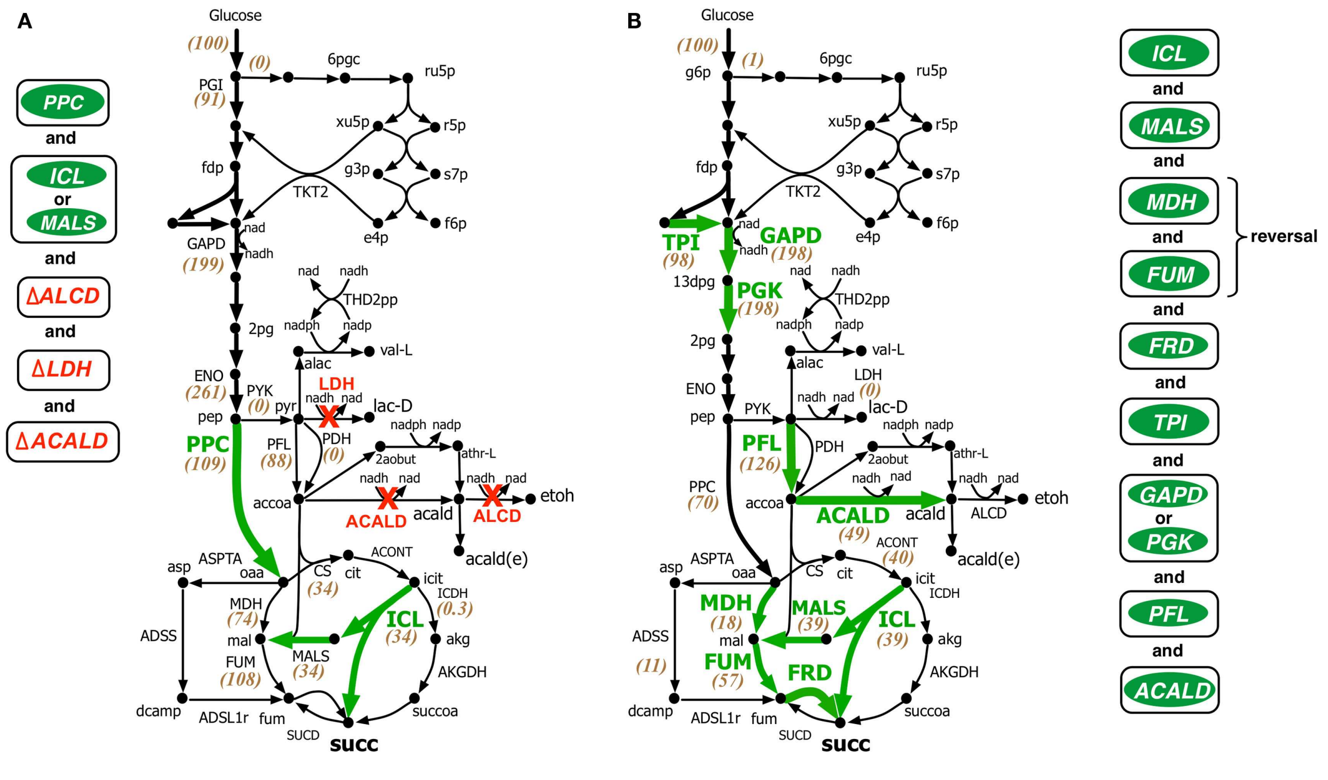


FIGURE 4 | Comparison of intervention strategies predicted by (A) regular OptForce and (B) k-OptForce for over production of succinate under anaerobic condition in *E. coli*. The values within parentheses indicate the metabolic flux in mmol gDW^{-1} h^{-1} per 100 mmol gDW^{-1} h^{-1} glucose uptake.

The kinetic model also fails to capture the metabolic transition of *E. coli* central metabolism from aerobic to anaerobic condition due to lack of regulatory information (Salmon et al., 2003, 2005). Under anaerobic condition, PP pathway, PPC, and TCA cycle are repressed, while glycolysis and, in particular, fermentative pathways are up-regulated (Perrenoud and Sauer, 2005; Cho et al., 2006). In addition, pyruvate dehydrogenase (PDH) is deactivated while PFL carries most of the flux from pyruvate to acetyl-CoA (Partridge et al., 2006). Even though the kinetic model captures down-regulation of TCA cycle upon removal of oxygen it cannot capture the remaining changes. Unable to capture the repression of PPC [anaerobic PPC flux is one-tenth of aerobic flux (Choudhary et al., 2011)], k-OptForce does not suggest any up-regulation in its activity to push more flux from pep towards oaa, contrary to OptForce suggestion of a minimum 15-fold up-regulation in PPC flux (8.4–133.3 mmol $gDW^{-1} h^{-1}$). In contrast, failing to recognize the regulatory activation of PFL under anaerobic condition, k-OptForce suggests a minimum eightfold up-regulation in its activity, while OptForce requires no such intervention. Unable to recognize the up-regulation of the enzyme activities in the fermentative pathways in the reference (non-engineered) strain, k-OptForce does not suggest any down-regulations since the parameterization of the enzymes does not allow a significant amount of flux towards ethanol, acetate, and lactate. In contrast, OptForce requires the removal of lactate dehydrogenase (LDH), alcohol dehydrogenase (ALCD), and acetaldehyde dehydrogenase (ACALD) to prevent diverting pyruvate flux away from succinate. Surprisingly, k-OptForce suggests a fivefold up-regulation in ACALD activity to maintain NAD/NADH redox balance. A large fraction of the produced acetaldehyde is reduced to ethanol (46 mmol $gDW^{-1} h^{-1}$), while the rest is exported out of the cell (3 mmol $gDW^{-1} h^{-1}$). However, we note that as no information capturing the effect of acetaldehyde on cell fitness was included in the kinetic model, it is unable to capture the chemical's toxicity. k-OptForce also suggests a minimum 1.5-fold up-regulation in triose phosphate isomerase (TPI) activity and a twofold up-regulation in GAPD or phosphoglycerate kinase (PGK) activity to route additional PP pathway flux through glycolysis, even though the PP pathway is negligibly active in anaerobic condition (Choudhary et al., 2011). It is to be noted here that down-regulation of TKT2 for aerobic overproduction of succinate and up-regulation of GAPD for anaerobic case are not equivalent interventions even though both strategies do increase glycolytic flux. This is because, the flux distribution in the pay-off phase of glycolysis, which is different in both cases, affects the metabolite concentrations of the preparatory phase of glycolysis. Up-regulation of ENO in aerobic overproduction study pulls additional metabolic flux down from upper glycolysis in addition to TKT2 removal. In absence of ENO up-regulation, removal of TKT2 cannot reroute the entire amount of PP flux towards glycolysis. As a result, up-regulation of both GAPD and PGK (and TPI) is necessary. It is also to be noted that the inactivation of PDH (and the subsequent activation of PFL) in anaerobic condition affects the reactions preceding it.

Comparison with experimental studies shows that unlike in the aerobic case, most of the verified engineering strategies are consistent with OptForce suggestions. k-OptForce overlooks key interventions such as up-regulation of PPC and removal of fermentative pathways, that were identified to have the largest impact in enhancing succinate yield (Millard et al., 1996; Zhang et al., 2009). In addition, even in cases where k-OptForce correctly identifies interventions, such as of MDH, FUM, and FRD up-regulation, inaccurate parameterization result in yield predictions far below experimentally observed succinate yield [1.08 vs. 1.2–1.6 mol/mol glucose with fewer interventions (Cao et al., 2013)]. In other cases, untested interventions such as up-regulation of PFL most likely will not improve succinate yield, considering that the deletion of *pflB* was found to improve succinate yield (Sanchez et al., 2005; Wu et al., 2007).

DISCUSSION

In this study, we compared the performance of k-OptForce in predicting interventions for overproduction of succinate in *E. coli* under both aerobic and anaerobic conditions. k-OptForce predictions under aerobic condition was found to be much more consistent with experimental strain-design strategies as compared with the stoichiometry-only OptForce predictions. In contrast, interventions for succinate overproduction under anaerobic condition by k-OptForce led to significantly less promising strategies largely inconsistent with experimental observations. This indicates that kinetic models have the potential to substantially over-perform FBA predictions when parameterized under the same (or similar) conditions but they may perform worse than FBA when asked to predict a significantly different metabolic phenotype. Note that the two-step strategy of the k-OptForce procedure does not affect the optimality of the results for the aerobic case as all interventions were identified from the kinetic part of the model. The flux distribution in the stoichiometric part of the model, which is determined by the worst-case inner problem, was effectively locked by the kinetic expressions. In general, however, we may miss better intervention strategies (for example in the anaerobic case study) when implementing the two-step approach as a tradeoff for improving computational performance.

The kinetic model was successful in capturing the underlying kinetic regulation when the flux re-distribution was consistent with the mutant flux information used for parameterizing the kinetic model. For example, the effect of enzymatic interventions around glycolysis and TCA cycle were identified with reasonable accuracy in both anaerobic and aerobic cases. Under aerobic condition, the kinetic model successfully captures the need for equimolar amounts of acetyl-CoA and oaa to supply the TCA cycle while preventing accumulation of intermediates (Lin et al., 2005a). Even when the kinetic model failed to correctly quantify fluxes, it provided a qualitative basis for making the right interventions. For example, k-OptForce correctly identifies that up-regulation of MDH, FUM, and FRD improves succinate production under anaerobic condition, even though it over-estimates the kinetic bottleneck towards such a flux-reversal resulting in poorer yields than experimental observations. Note that the developed kinetic model cannot capture changes in glucose uptake rate for different environmental and/or genetic backgrounds as all mutant fluxes used to train the model were scaled with the corresponding glucose uptake. Shortcomings in the model could be rectified by re-parameterizing the model using additional fluxomic information of mutant strains that allow for pathway reversal [e.g., using metabolic flux analysis information of a ΔSUCD strain (Li et al., 2006)]. In general, the re-parameterization is a compromise between model scope and accuracy. The observations showed that parameterizing the kinetic model by making use of mutant data located in the proximity of a target product provides a more accurate flux distribution predictions by the model and consequently results to the identification of more targeted interventions using the k-OptForce procedure. In contrast, integration of a wide-range of conditions with limited experimental data for model training may provide a better global qualitative agreement. While one could use separate kinetic models for aerobic and anaerobic conditions, ideally we would like a single model parameterization that could reproduce both aerobic and anaerobic responses. By creating two separate aerobic and anaerobic models it becomes unclear what model to use under micro/partial aerobic condition (Partridge et al., 2007).

This study shows that the model does not retain fidelity of predictions when growth is switched from aerobic to anaerobic condition. Aerobic to anaerobic metabolic transition is mainly controlled at the transcriptional level (Kochanowski et al., 2013) by the activities of global regulatory proteins FNR and ArcA (see **Table 2**). In absence of such regulatory interactions, the kinetic model could not capture the activation of PFL and fermentative pathways, and the deactivation of PPC and (to a small extent) PP Pathway. As a result, k-OptForce failed to identify key down-regulations (e.g., LDH, ALCD) in the former case, while suggested unnecessary up-regulations for the latter. These shortcomings are harder to address and require the incorporation of adequate regulatory information into the model (see **Table 2** for details) to capture the aerobic to anaerobic transition.

Table 2 | Regulatory systems under anaerobic condition in *E. coli* (Partridge et al., 2006).

Regulator	Type	Target gene	Target reaction
ArcA	Repression	*sucABCD*	SUCOAS
		sdhABCD	SUCD
		fumA	FUM
		mdh	MDH
		aceEF	PDH
		acnAB	ACONT
		gltA	CS
		icdA	ICDH
	Activation	*pfl*	PFL
FNR	Repression	*acnA*	ACONT
		icdA	ICDH
		sdhABCD	SUCD
		fumAC	FUM
		ndh	NDH

SUCOAS, succinyl-CoA synthetase; SUCD, succinate dehydrogenase; FUM, fumarase; MDH, malate dehydrogenase; PDH, pyruvate dehydrogenase; ACONT, aconitase; CS, citrate synthase; ICDH, isocitrate dehydrogenase; PFL, pyruvate formate lyase; NDH, nadh dehydrogenase.

In general, this study revealed some of the strengths and limitations of kinetic model-driven strain design. It demonstrated the need to carry out model parameterization for a diverse range of genetic/environmental perturbations (Khodayari et al., 2014) and the tight integration of transcriptional level along with substrate-level regulatory interactions. At a fundamental level, kinetic models must be *a priori* provided with the quantitative description of as many as possible regulatory switches that become active in response to genetic or environmental perturbations. This richness in mechanistic information enables a detailed description of metabolism that captures dynamics, enzyme activities, and metabolite concentrations but can lead to erroneous predictions due to missing and/or incorrect modeling assumptions. Nevertheless, by studying failure modes of kinetic models, valuable information can be uncovered for restoring prediction consistency for new phenotypes.

AUTHOR CONTRIBUTIONS

Conceived and designed experiments: Costas D. Maranas, Anupam Chowdhury, Ali Khodayari. Performed the experiments: Anupam Chowdhury and Ali Khodayari. Analyzed the data: Anupam Chowdhury, Ali Khodayari, Costas D. Maranas. Contributed reagents/materials/analysis tools: Anupam Chowdhury, Ali Khodayari, Costas D. Maranas. Wrote paper: Ali Khodayari, Anupam Chowdhury, Costas D. Maranas.

ACKNOWLEDGMENTS

The authors gratefully acknowledge funding from the NSF (http://www.nsf.gov/) award no. EEC-0813570 and the DOE (http://www.energy.gov/) grant no. DE-SC10822882. The funders had no role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript.

REFERENCES

Almquist, J., Cvijovic, M., Hatzimanikatis, V., Nielsen, J., and Jirstrand, M. (2014). Kinetic models in industrial biotechnology – improving cell factory performance. *Metab. Eng.* 24, 38–60. doi:10.1016/j.ymben.2014.03.007

Angermayr, S. A., and Hellingwerf, K. J. (2013). On the use of metabolic control analysis in the optimization of cyanobacterial biosolar cell factories. *J. Phys. Chem. B* 117, 11169–11175. doi:10.1021/jp4013152

Baez-Viveros, J. L., Flores, N., Juarez, K., Castillo-Espana, P., Bolivar, F., and Gosset, G. (2007). Metabolic transcription analysis of engineered *Escherichia coli* strains that overproduce L-phenylalanine. *Microb. Cell Fact.* 6, 30. doi:10.1186/1475-2859-6-30

Burgard, A. P., Pharkya, P., and Maranas, C. D. (2003). Optknock: a bilevel programming framework for identifying gene knockout strategies for microbial strain optimization. *Biotechnol. Bioeng.* 84, 647–657. doi:10.1002/bit.10803

Cao, Y., Cao, Y., and Lin, X. (2011). Metabolically engineered *Escherichia coli* for biotechnological production of four-carbon 1,4-dicarboxylic acids. *J. Ind. Microbiol. Biotechnol.* 38, 649–656. doi:10.1007/s10295-010-0913-4

Cao, Y., Zhang, R., Sun, C., Cheng, T., Liu, Y., and Xian, M. (2013). Fermentative succinate production: an emerging technology to replace the traditional petrochemical processes. *Biomed Res. Int.* 2013, 723412. doi:10.1155/2013/723412

Cho, B. K., Knight, E. M., and Palsson, B. O. (2006). Transcriptional regulation of the fad regulon genes of *Escherichia coli* by ArcA. *Microbiology* 152, 2207–2219. doi:10.1099/mic.0.28912-0

Choudhary, M. K., Yoon, J. M., Gonzalez, R., and Shanks, J. V. (2011). Re-examination of metabolic fluxes in *Escherichia coli* during anaerobic fermentation of glucose using C-13 labeling experiments and 2-dimensional nuclear magnetic resonance (NMR) spectroscopy. *Biotechnol. Bioprocess Eng.* 16, 419–437. doi:10.1007/s12257-010-0449-5

Chowdhury, A., Zomorrodi, A. R., and Maranas, C. D. (2014). k-OptForce: integrating kinetics with flux balance analysis for strain design. *PLoS Comput. Biol.* 10:e1003487. doi:10.1371/journal.pcbi.1003487

Cotten, C., and Reed, J. L. (2013a). Constraint-based strain design using continuous modifications (CosMos) of flux bounds finds new strategies for metabolic engineering. *Biotechnol. J.* 8, 595–604. doi:10.1002/biot.201200316

Cotten, C., and Reed, J. L. (2013b). Mechanistic analysis of multi-omics datasets to generate kinetic parameters for constraint-based metabolic models. *BMC Bioinformatics* 14:32. doi:10.1186/1471-2105-14-32

Curran, K. A., and Alper, H. S. (2012). Expanding the chemical palate of cells by combining systems biology and metabolic engineering. *Metab. Eng.* 14, 289–297. doi:10.1016/j.ymben.2012.04.006

Duckworth, H. W., Barber, B. H., and Sanwal, B. D. (1973). The interaction of phosphoglucomutase with nucleotide inhibitors. *J. Biol. Chem.* 248, 1431–1435.

Feist, A. M., Henry, C. S., Reed, J. L., Krummenacker, M., Joyce, A. R., Karp, P. D., et al. (2007). A genome-scale metabolic reconstruction for *Escherichia coli* K-12 MG1655 that accounts for 1260 ORFs and thermodynamic information. *Mol. Syst. Biol.* 3, 121. doi:10.1038/msb4100155

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

Fleming, R. M., Thiele, I., Provan, G., and Nasheuer, H. P. (2010). Integrated stoichiometric, thermodynamic and kinetic modelling of steady state metabolism. *J. Theor. Biol.* 264, 683–692. doi:10.1016/j.jtbi.2010.02.044

Flowers, D., Thompson, R. A., Birdwell, D., Wang, T., and Trinh, C. T. (2013). SMET: systematic multiple enzyme targeting – a method to rationally design optimal strains for target chemical overproduction. *Biotechnol. J.* 8, 605–618. doi:10.1002/biot.201200233

Grant, G. A. (2012). Contrasting catalytic and allosteric mechanisms for phosphoglycerate dehydrogenases. *Arch. Biochem. Biophys.* 519, 175–185. doi:10.1016/j.abb.2011.10.005

Grossmann, I. E., Viswanathan, J., Vecchietti, A., Raman, R., and Kalvelagen, E. (2002). *GAMS/DICOPT: A Discrete Continuous Optimization Package.* Washington, DC: GAMS Development Corporation.

Gruys, K. J., Walker, M. C., and Sikorski, J. A. (1992). Substrate synergism and the steady-state kinetic reaction mechanism for EPSP synthase from *Escherichia coli. Biochemistry* 31, 5534–5544. doi:10.1021/bi00139a016

Heijnen, J. J., and Verheijen, P. J. (2013). Parameter identification of in vivo kinetic models: limitations and challenges. *Biotechnol. J.* 8, 768–775. doi:10.1002/biot.201300105

Hong, K. K., and Nielsen, J. (2012). Metabolic engineering of *Saccharomyces cerevisiae*: a key cell factory platform for future biorefineries. *Cell. Mol. Life Sci.* 69, 2671–2690. doi:10.1007/s00018-012-0945-1

Hoyt, J. C., Robertson, E. F., Berlyn, K. A., and Reeves, H. C. (1988). *Escherichia coli* isocitrate lyase: properties and comparisons. *Biochim. Biophys. Acta* 966, 30–35. doi:10.1016/0304-4165(88)90125-0

Hubbard, B. K., Koch, M., Palmer, D. R., Babbitt, P. C., and Gerlt, J. A. (1998). Evolution of enzymatic activities in the enolase superfamily: characterization of the (D)-glucarate/galactarate catabolic pathway in *Escherichia coli. Biochemistry* 37, 14369–14375. doi:10.1021/bi981124f

Ishii, N., Nakahigashi, K., Baba, T., Robert, M., Soga, T., Kanai, A., et al. (2007). Multiple high-throughput analyses monitor the response of *E. coli* to perturbations. *Science* 316, 593–597. doi:10.1126/science.1132067

Jamshidi, N., and Palsson, B. O. (2010). Mass action stoichiometric simulation models: incorporating kinetics and regulation into stoichiometric models. *Biophys. J.* 98, 175–185. doi:10.1016/j.bpj.2009.09.064

Jouhten, P. (2012). Metabolic modelling in the development of cell factories by synthetic biology. *Comput. Struct. Biotechnol. J.* 3, 9. doi:10.5936/csbj.201210009

Khodayari, A., Zomorrodi, A. R., Liao, J. C., and Maranas, C. D. (2014). A kinetic model of *Escherichia coli* core metabolism satisfying multiple sets of mutant flux data. *Metab. Eng.* 25C, 50–62. doi:10.1016/j.ymben.2014.05.014

Kim, J., and Reed, J. L. (2010). OptORF: optimal metabolic and regulatory perturbations for metabolic engineering of microbial strains. *BMC Syst. Biol.* 4:53. doi:10.1186/1752-0509-4-53

Kim, J., Reed, J. L., and Maravelias, C. T. (2011). Large-scale bi-level strain design approaches and mixed-integer programming solution techniques. *PLoS ONE* 6:e24162. doi:10.1371/journal.pone.0024162

Kochanowski, K., Sauer, U., and Chubukov, V. (2013). Somewhat in control – the role of transcription in regulating microbial metabolic fluxes. *Curr. Opin. Biotechnol.* 24, 987–993. doi:10.1016/j.copbio.2013.03.014

Lai, S., Zhang, Y., Liu, S., Liang, Y., Shang, X., Chai, X., et al. (2012). Metabolic engineering and flux analysis of *Corynebacterium glutamicum* for L-serine production. *Sci. China Life Sci.* 55, 283–290. doi:10.1007/s11427-012-4304-0

Lee, J. W., Na, D., Park, J. M., Lee, J., Choi, S., and Lee, S. Y. (2012). Systems metabolic engineering of microorganisms for natural and non-natural chemicals. *Nat. Chem. Biol.* 8, 536–546. doi:10.1038/nchembio.970

Lee, K. H., Park, J. H., Kim, T. Y., Kim, H. U., and Lee, S. Y. (2007). Systems metabolic engineering of *Escherichia coli* for L-threonine production. *Mol. Syst. Biol.* 3, 149. doi:10.1038/msb4100196

Lee, S. J., Lee, D. Y., Kim, T. Y., Kim, B. H., Lee, J., and Lee, S. Y. (2005). Metabolic engineering of *Escherichia coli* for enhanced production of succinic acid, based on genome comparison and in silico gene knockout simulation. *Appl. Environ. Microbiol.* 71, 7880–7887. doi:10.1128/AEM.71.12.7880-7887.2005

Li, M., Ho, P. Y., Yao, S., and Shimizu, K. (2006). Effect of sucA or sucC gene knockout on the metabolism in *Escherichia coli* based on gene expressions, enzyme activities, intracellular metabolite concentrations and metabolic fluxes by 13C-labeling experiments. *Biochem. Eng. J.* 30, 289–296. doi:10.1016/j.bej.2006.05.011

Li, Y., Chen, G. K., Tong, X. W., Zhang, H. T., Liu, X. G., Liu, Y. H., et al. (2012). Construction of *Escherichia coli* strains producing L-serine from glucose. *Biotechnol. Lett.* 34, 1525–1530. doi:10.1007/s10529-012-0937-0

Lin, H., Bennett, G. N., and San, K. Y. (2005a). Chemostat culture characterization of *Escherichia coli* mutant strains metabolically engineered for aerobic succinate production: a study of the modified metabolic network based on metabolite profile, enzyme activity, and gene expression profile. *Metab. Eng.* 7, 337–352. doi:10.1016/j.ymben.2005.06.002

Lin, H., Bennett, G. N., and San, K. Y. (2005b). Fed-batch culture of a metabolically engineered *Escherichia coli* strain designed for high-level succinate production and yield under aerobic conditions. *Biotechnol. Bioeng.* 90, 775–779. doi:10.1002/bit.20458

Litsanov, B., Kabus, A., Brocker, M., and Bott, M. (2012). Efficient aerobic succinate production from glucose in minimal medium with *Corynebacterium glutamicum. Microb. Biotechnol.* 5, 116–128. doi:10.1111/j.1751-7915.2011.00310.x

MacKintosh, C., and Nimmo, H. G. (1988). Purification and regulatory properties of isocitrate lyase from *Escherichia coli* ML308. *Biochem. J.* 250, 25–31.

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

Maia, P., Vilaca, P., Rocha, I., Pont, M., Tomb, J. F., and Rocha, M. (2012). An integrated computational environment for elementary modes analysis of biochemical networks. *Int. J. Data Min. Bioinform.* 6, 382–395. doi:10.1504/IJDMB.2012.049292

Millard, C. S., Chao, Y. P., Liao, J. C., and Donnelly, M. I. (1996). Enhanced production of succinic acid by overexpression of phosphoenolpyruvate carboxylase in *Escherichia coli. Appl. Environ. Microbiol.* 62, 1808–1810.

Nikolaev, E. V. (2010). The elucidation of metabolic pathways and their improvements using stable optimization of large-scale kinetic models of cellular systems. *Metab. Eng.* 12, 26–38. doi:10.1016/j.ymben.2009.08.010

Ogawa, T., Murakami, K., Mori, H., Ishii, N., Tomita, M., and Yoshin, M. (2007). Role of phosphoenolpyruvate in the NADP-isocitrate dehydrogenase and isocitrate lyase reaction in *Escherichia coli. J. Bacteriol.* 189, 1176–1178. doi:10.1128/JB.01628-06

Osterhout, R. E., Pharkya, P., and Burgard, A. P. (2011). Microorganisms and methods for the production of ethylene glycol. US 13/086,295.

Partridge, J. D., Sanguinetti, G., Dibden, D. P., Roberts, R. E., Poole, R. K., and Green, J. (2007). Transition of *Escherichia coli* from aerobic to micro-aerobic conditions involves fast and slow reacting regulatory components. *J. Biol. Chem.* 282, 11230–11237. doi:10.1074/jbc.M700728200

Partridge, J. D., Scott, C., Tang, Y., Poole, R. K., and Green, J. (2006). *Escherichia coli* transcriptome dynamics during the transition from anaerobic to aerobic conditions. *J. Biol. Chem.* 281, 27806–27815. doi:10.1074/jbc.M603450200

Perrenoud, A., and Sauer, U. (2005). Impact of global transcriptional regulation by ArcA, ArcB, Cra, Crp, Cya, Fnr, and Mlc on glucose catabolism in *Escherichia coli. J. Bacteriol.* 187, 3171–3179. doi:10.1128/JB.187.9.3171-3179.2005

Pharkya, P., Burgard, A. P., and Maranas, C. D. (2004). OptStrain: a computational framework for redesign of microbial production systems. *Genome Res.* 14, 2367–2376. doi:10.1101/gr.2872004

Ranganathan, S., Suthers, P. F., and Maranas, C. D. (2010). OptForce: an optimization procedure for identifying all genetic manipulations leading to targeted overproductions. *PLoS Comput. Biol.* 6:e1000744. doi:10.1371/journal.pcbi.1000744

Rocha, I., Maia, P., Evangelista, P., Vilaca, P., Soares, S., Pinto, J. P., et al. (2010). OptFlux: an open-source software platform for in silico metabolic engineering. *BMC Syst. Biol.* 4:45. doi:10.1186/1752-0509-4-45

Sahinidis, N. V. (1996). BARON: a general purpose global optimization software package. *J. Global Optim.* 8, 201–205. doi:10.1007/BF00138693

Salmon, K., Hung, S. P., Mekjian, K., Baldi, P., Hatfield, G. W., and Gunsalus, R. P. (2003). Global gene expression profiling in *Escherichia coli* K12. The effects of oxygen availability and FNR. *J. Biol. Chem.* 278, 29837–29855. doi:10.1074/jbc.M213060200

Salmon, K. A., Hung, S. P., Steffen, N. R., Krupp, R., Baldi, P., Hatfield, G. W., et al. (2005). Global gene expression profiling in *Escherichia coli* K12: effects of oxygen availability and ArcA. *J. Biol. Chem.* 280, 15084–15096. doi:10.1074/jbc.M414030200

Sanchez, A. M., Bennett, G. N., and San, K. Y. (2005). Efficient succinic acid production from glucose through overexpression of pyruvate carboxylase in an *Escherichia coli* alcohol dehydrogenase and lactate dehydrogenase mutant. *Biotechnol. Prog.* 21, 358–365. doi:10.1021/bp049676e

Sanwal, B. D., Duckworth, H. W., and Hollier, M. L. (1972). Regulation of phosphoglucomutase. *Biochem. J.* 128, 26–27.

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

Smallbone, K., Simeonidis, E., Swainston, N., and Mendes, P. (2010). Towards a genome-scale kinetic model of cellular metabolism. *BMC Syst. Biol.* 4:6. doi:10.1186/1752-0509-4-6

Song, H. S., and Ramkrishna, D. (2012). Prediction of dynamic behavior of mutant strains from limited wild-type data. *Metab. Eng.* 14, 69–80. doi:10.1016/j.ymben.2012.02.003

Sprenger, G. A., Schorken, U., Sprenger, G., and Sahm, H. (1995). Transaldolase B of *Escherichia coli* K-12: cloning of its gene, talB, and characterization of the enzyme from recombinant strains. *J. Bacteriol.* 177, 5930–5936.

Tan, Y., and Liao, J. C. (2012). Metabolic ensemble modeling for strain engineers. *Biotechnol. J.* 7, 343–353. doi:10.1002/biot.201100186

Tan, Y., Rivera, J. G., Contador, C. A., Asenjo, J. A., and Liao, J. C. (2011). Reducing the allowable kinetic space by constructing ensemble of dynamic models with the same steady-state flux. *Metab. Eng.* 13, 60–75. doi:10.1016/j.ymben.2010.11.001

Tepper, N., and Shlomi, T. (2010). Predicting metabolic engineering knockout strategies for chemical production: accounting for competing pathways. *Bioinformatics* 26, 536–543. doi:10.1093/bioinformatics/btp704

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

Villaverde, F. A., Henriques, D., Smallbone, K., Bongard, S., Schmid, J., Cicin-Sain, D., et al. (2014). BioPreDyn-bench: benchmark problems for kinetic modelling in systems biology. *BioPreDyn-Bench* arXiv:1407.5856.

Wang, Q., Qi, Y., Yin, N., and Lai, L. (2014). Discovery of novel allosteric effectors based on the predicted allosteric sites for *Escherichia coli* D-3-phosphoglycerate dehydrogenase. *PLoS ONE* 9:e94829. doi:10.1371/journal.pone.0094829

Wu, H., Li, Z. M., Zhou, L., and Ye, Q. (2007). Improved succinic acid production in the anaerobic culture of an *Escherichia coli* pflB ldhA double mutant as a result of enhanced anaplerotic activities in the preceding aerobic culture. *Appl. Environ. Microbiol.* 73, 7837–7843. doi:10.1128/AEM.01546-07

Xu, P., Ranganathan, S., Fowler, Z. L., Maranas, C. D., and Koffas, M. A. (2011). Genome-scale metabolic network modeling results in minimal interventions that cooperatively force carbon flux towards malonyl-CoA. *Metab. Eng.* 13, 578–587. doi:10.1016/j.ymben.2011.06.008

Zhang, X., Jantama, K., Moore, J. C., Jarboe, L. R., Shanmugam, K. T., and Ingram, L. O. (2009). Metabolic evolution of energy-conserving pathways for succinate production in *Escherichia coli*. *Proc. Natl. Acad. Sci. U.S.A.* 106, 20180–20185. doi:10.1073/pnas.0905396106

Zhu, N., Xia, H., Wang, Z., Zhao, X., and Chen, T. (2013). Engineering of acetate recycling and citrate synthase to improve aerobic succinate production in *Corynebacterium glutamicum*. *PLoS ONE* 8:e60659. doi:10.1371/journal.pone.0060659

Zomorrodi, A. R., Suthers, P. F., Ranganathan, S., and Maranas, C. D. (2012). Mathematical optimization applications in metabolic networks. *Metab. Eng.* 14, 672–686. doi:10.1016/j.ymben.2012.09.005

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.

Received: 03 September 2014; accepted: 05 December 2014; published online: 05 January 2015.

Citation: Khodayari A, Chowdhury A and Maranas CD (2015) Succinate overproduction: a case study of computational strain design using a comprehensive Escherichia coli kinetic model. Front. Bioeng. Biotechnol. ***2****:76. doi: 10.3389/fbioe.2014.00076*

This article was submitted to Systems Biology, a section of the journal Frontiers in Bioengineering and Biotechnology.

REVIEW ARTICLE
published: 08 December 2014
doi: 10.3389/fbioe.2014.00061

Improving collaboration by standardization efforts in systems biology

Andreas Dräger[1,2]* and *Bernhard Ø. Palsson*[1]

[1] *Systems Biology Research Group, Department of Bioengineering, University of California, San Diego, La Jolla, CA, USA*
[2] *Cognitive Systems, Center for Bioinformatics Tübingen (ZBIT), Department of Computer Science, University of Tübingen, Tübingen, Germany*

Edited by:
Daniel Machado, University of Minho, Portugal

Reviewed by:
Padraig Gleeson, University College London, UK
Taishin Nomura, Osaka University, Japan
Herbert M. Sauro, University of Washington, USA

Correspondence:
Andreas Dräger, University of California, Atkinson Hall, Room 2506, 9500 Gilman Drive # 0412, La Jolla, CA 92093-0412, USA
e-mail: andraeger@eng.ucsd.edu

Collaborative genome-scale reconstruction endeavors of metabolic networks would not be possible without a common, standardized formal representation of these systems. The ability to precisely define biological building blocks together with their dynamic behavior has even been considered a prerequisite for upcoming synthetic biology approaches. Driven by the requirements of such ambitious research goals, standardization itself has become an active field of research on nearly all levels of granularity in biology. In addition to the originally envisaged exchange of computational models and tool interoperability, new standards have been suggested for an unambiguous graphical display of biological phenomena, to annotate, archive, as well as to rank models, and to describe execution and the outcomes of simulation experiments. The spectrum now even covers the interaction of entire neurons in the brain, three-dimensional motions, and the description of pharmacometric studies. Thereby, the mathematical description of systems and approaches for their (repeated) simulation are clearly separated from each other and also from their graphical representation. Minimum information definitions constitute guidelines and common operation protocols in order to ensure reproducibility of findings and a unified knowledge representation. Central database infrastructures have been established that provide the scientific community with persistent links from model annotations to online resources. A rich variety of open-source software tools thrives for all data formats, often supporting a multitude of programing languages. Regular meetings and workshops of developers and users lead to continuous improvement and ongoing development of these standardization efforts. This article gives a brief overview about the current state of the growing number of operation protocols, mark-up languages, graphical descriptions, and fundamental software support with relevance to systems biology.

Keywords: model formats, modeling guidelines, ontologies, model databases, network visualization, software support

1. INTRODUCTION

Since its emergence in the 1960s systems biology has always been tightly related to the availability of powerful computational resources. While at the beginning of research in the field and its applications quick and simple script-based solutions were sufficient, the bar for publication and review has been drastically raised (Sauro et al., 2003). It has been realized that individual scripts, which are specific to certain computational environments and that are not very reproducible are of small benefit for the scientific community and progress of the field (Lloyd et al., 2004). The development of standardized data formats, models, and computational methods have paved the way toward the evolution and maturation of systems biology into a main-stream field of research (Macilwain, 2011). Sufficient annotation and metadata of models, experiments, and other data enhance the reproducibility of

Abbreviations: ANSI, American National Standards Institute; API, application programing interface; BRAIN, brain research through advancing innovative neurotechnologies; CAD, computer-aided design; COPASI, complex pathway simulator; CSS, cascading style sheets; DAE, differential-algebraic equation; DIN, Deutsches Institut für Normung; FBA, flux balance analysis; fbc, flux balance constraints; GO, gene ontology; HTML, hyper text mark-up language; IEEE, Institute of Electrical and Electronics Engineers; IETF, internet engineering task force; ISML, *in silico* mark-up language; JSON, JavaScript object notation; KiSAO, kinetic simulation algorithm ontology; LEMS, low entropy model specification; MAMO, mathematical modeling ontology; MIASE, minimum information about a simulation experiment; MIBBI, minimal information for biological and biomedical research; MIRIAM, minimal information required in the annotation of models; NCBI, National Center for Biotechnology Information; NuML, numerical mark-up language; OBO, open biomedical ontologies; ODE, ordinary differential equation; OMEX, open modeling exchange format; OMG, object management group; OSB, open-source brain; OWL, web ontology language; PDE, partial differential equation; PharmML, pharmacometrics mark-up language; PHML, physiological hierarchy mark-up language; RDF, resource description framework; SBGN, systems biology graphical notation; SBGN-ML, systems biology graphical notation mark-up language; SBML, systems biology mark-up language; SBOL, synthetic biology open language; SBRML, systems biology result mark-up language; SBW, systems biology workbench; SED-ML, simulation experiment description mark-up language; SVG, scalable vector graphics; SWIG, simplified wrapper and interface generator; TEDDY, terminology for the description of dynamics; URI, uniform resource identifier; W3C, world wide web consortium; XML, eXtended mark-up language.

results (Wolstencroft et al., 2011). For individual areas of research, different models are required, hence different standards for their encoding. Research in constraint-based modeling (Bordbar et al., 2014) deals with the encoding of the stoichiometric matrix and flux bounds, whereas, dynamic metabolic modeling (Dräger and Planatscher, 2013a) is usually based on building ordinary differential equation systems, model calibration, and parameter estimation (Dräger et al., 2009a; Kronfeld et al., 2009; Dräger and Planatscher, 2013b). Spatial-temporal simulations require encoding three-dimensional geometries and partial differential equation systems (Moraru et al., 2008).

It can hence be observed that the modeling community in systems biology has diversified. One reason for this development is that main parts of funding for these standardization attempts originate from ambitious large-scale projects, each having has different requirements. These efforts include, for example, goal of specifically reconstructing all reactions in specific organisms, such as human or yeast, resulting in giant reaction networks (Duarte et al., 2007; Herrgård et al., 2008; Rolfsson et al., 2011; Thiele et al., 2013) or systematically representing the complete knowledge about biochemical reactions available today (Büchel et al., 2013a). Trans-European projects like SysMO (Booth, 2007) want to comprehensively record and describe dynamic molecular processes in unicellular microorganisms and to present all processes in the form of computerized mathematical models. The German Virtual Liver Network (Holzhütter et al., 2012) aims to mathematically explain all phenomena in the human liver across multiple cell types and levels of organization. The Physiome project attempts to achieve a full quantitative description of all physiological dynamics and functional behaviors of the intact human body (Hunter and Borg, 2003). The US BRAIN (Brain Research through Advancing Innovative Neurotechnologies) Initiative aims to support the development of new technologies for classifying the anatomical constituents for the brain and to allow simultaneous recording from an unprecedented number of neurons simultaneously. The EUs Human Brain Project seeks to develop the infrastructure for creating computational models of brain regions at multiple scales on high-performance computing platforms (Shepherd et al., 1998; Markram et al., 2011; Kandel et al., 2013). Thereby, medical applications become increasingly important (Büchel et al., 2013b; Grillner, 2014).

Common to all these consortia is that with the increasing number of active researchers and collaborators the exchange, reproduction, and accessibility of models, data, and further information in specific online databases play a major role (Brazma et al., 2006; Schellenberger et al., 2010; Wolstencroft et al., 2011; Yu et al., 2011; Chelliah et al., 2013). Just like the documentation of source code, the careful annotation of models and data are also necessary to achieve a fruitful collaboration. The more meta information that is provided, the easier the model can be comprehended, modified, simulated, and analyzed (Waltemath et al., 2013). The use of standard formats is highly recommended for the publication of results even if not required by the prospective journal.

In addition, new fields and areas of application are emerging, for instance, pharmacometric models or synthetic biology (Endler et al., 2009; Galdzicki et al., 2011; Müller and Arndt, 2012). There is therefore no one-size-fits-all solution that would be equally suitable for all fields of research. The standardization community therefore needs to continuously catch up with these developments in the actual modeling community and to reinvent itself over and over again. Recent approaches have suggested to modularize modeling languages by introducing highly specialized *packages* for modeling aspects that can otherwise not be represented in the main data format (Chaouiya et al., 2013).

The structure of how standards are defined has also matured. Brazma et al. (2006) describe that four steps are required for the development of a standard: (i) data and information need to be collected about the domain of interest that are relevant for an unambiguous transfer and interpretation as well as conceptual model design, (ii) the model needs to be formalized, (iii) an exchange format must be defined, and (iv) software support must be implemented. Nearly all modeling formats described in this article now follow this suggestion and are based on a minimum information requirement description (Taylor et al., 2008). These documents define what kind of information has to be stored in a respective model in order to guarantee that the model can be reused and understood by other researchers. In this way, the information requirement and the corresponding modeling standard are decoupled, exchangeable, and independent. The minimum information requirement is usually complemented with a specific ontology, i.e., a hierarchical collection of field-specific terms and their definitions (Courtot et al., 2011). These terms can be associated to model components and descriptions. In addition, elaborate and persistent annotation frameworks have been developed, which allow the modeler to precisely express, what individual model components are and how they are to be understood (Juty et al., 2012, 2013). The development of standards, minimal information requirements, and ontologies needs to be orthogonal to existing respective standards. **Table 1** and **Figure 1** give an overview about the relationship amongst various standards discussed in this article.

The structural representation of the model [for instance, SBML by Hucka et al. (2004) or CellML by Cuellar et al. (2006)], its application and analysis [SED-ML by Waltemath et al. (2011b)], its (graphical) display [SBGN guidelines by Le Novère et al. (2009)], and features should be accurately discriminated and encoded in

Table 1 | Standards with relevance for modeling in systems biology.

	Model	**Procedures**	**Results**
Representation formats	BioPAX, CellML, NeuroML, PharmML, SBML (including extension packages), SBGN-ML, SBOL	SED-ML	NuML, SBRML
Graphical display	CellML visualization, SBGN, SBOL visual		
Minimal information requirements	MIRIAM	MIASE	
Mathematical semantics	SBO, MAMO	KiSAO	TEDDY
Biological semantics	MIRIAM	MIRIAM	MIRIAM

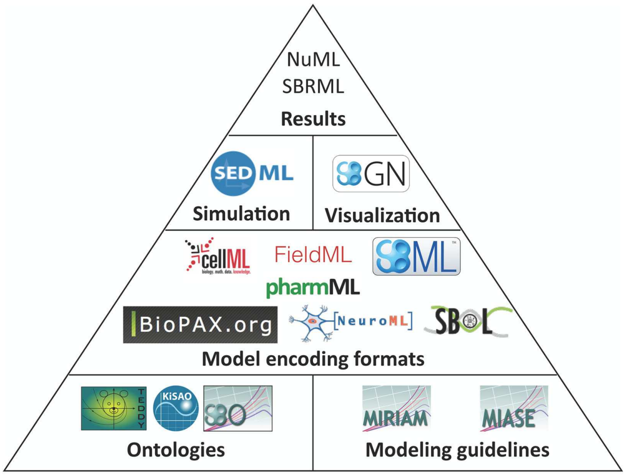


FIGURE 1 | Standards overview. Hierarchically organized controlled vocabularies, so-called *ontologies* and modeling guidelines build the basis for model encoding formats. These formats can refer to terms from ontologies and their organization is in accordance with the modeling guidelines. Recommendations for a visual representation of models as well as the execution of individual models in numerical simulation or optimization are separated from the structural models. Numerical results can be encoded in further standard data formats.

distinct formats. Depending on the concrete modeling format, structural models can also include mathematical formulations, but not their interpretation framework (such as the algorithm to solve the model or the simulation end time). Recently, a new archive format has been proposed in order to link and distribute these independent modeling aspects all together in a single file (Bergmann et al., 2014).

Much effort has also been invested in software support and the creation of infrastructures for diverse standards. For each data format, a specific library has been implemented for reading and writing files as well as for manipulating components of the format in memory (Bornstein et al., 2008; Miller et al., 2010; Demir et al., 2013). Often, language-bindings for diverse programing environments are provided, but sometimes specific libraries have been developed in order to support certain programing languages (Dräger et al., 2011). These parsing libraries help developers to use and exploit the individual standards. Often these libraries provide interfaces to corresponding ontologies and controlled vocabulary annotations (Courtot et al., 2011). However, the interpretation, analysis, drawing, etc. of models cannot be facilitated by these libraries. Higher level software has been implemented to support model building, display, simulation, etc. (Deckard et al., 2006; Keller et al., 2013). Sometimes, this is done in the form of plug-ins to more general frameworks, and often there are diverse stand-alone or web-based tools for various purposes (König et al., 2012; Krause et al., 2013).

When the first XML- or OWL-based exchange formats for models were proposed, developers of existing software tools were often involved, and their individual software was adapted in order to fit the standard. Nowadays, with many standards being well established, software is specifically tailored with respect to the standards. The stringent elaboration and clear distinction between models, purpose, simulation, and annotation can also be a source of inspiration for young researchers who enter the field. In the long-term, using standard formats can lower the expenses for software development because they allow the reuse of existing tools in new applications. Moreover, with the many available tools for standard formats, less research time is needed for the interconversion of tool-specific files, making it much easier to collect information from diverse sources (Demir et al., 2010).

While international and national standardization bodies, such as OMG, W3C, IEEE, ANSI, IETF, DIN, etc., usually approve standards and release specifications, the situation is different in systems biology, where *de facto* standards are established by the scientific community (Brazma et al., 2006). The fast-moving nature and ongoing development of research makes this approach necessary.

However, keeping track of the growing number of model formats and standards for diverse purposes has become more and more difficult. This review article gives a broad overview of a wide range of currently existing modeling standards, formats, and online repositories, and a selection of software solutions for systems biology and related fields of research. The aim of this article is to highlight specific standards, their usability, and application in order to give the reader an up-to-date picture of model definition, encoding, and availability in systems biology.

2. MATERIAL AND METHODS

2.1. MODELING GUIDELINES

Modeling formats give us the syntax of models (Juty et al., 2012). In order to enhance accessibility of data and to facilitate the reuse of models, several modeling guidelines have been proposed, which are discussed in this section. These guidelines are often called "Minimum Information of/for," which should express that without at least this form of information optimal use and reproducibility of results cannot be guaranteed. More information can always be provided on top of the minimal requirements. The guidelines are hence a form of checklists that describe which kind of information to include and often go back to the idea of the MIBBI project (Minimal Information for Biological and Biomedical research) proposed by Taylor et al. (2008). The open biomedical ontologies (OBO) foundry[1] maintains orthogonal (non-overlapping) collections of controlled vocabularies, which provide the semantics for models. The most well-known ontology is probably the gene ontology (GO) by (Ashburner et al., 2000).

2.1.1. Minimum information required in the annotation of models

Reuse of models can be compromised if inconsistent identifier systems are used for individual components. For instance, when merging models, it is necessary to match overlapping components. If a molecule is identified as water in one model and as H_2O in another such a matching is already difficult for automated procedures. To solve such problems, the minimum information required in the annotation of models (MIRIAM) guidelines has been proposed as a general model curation checklist (Le Novère et al., 2005). The MIRIAM registry (Laible and Le Novère, 2007) goes further

[1] http://obofoundry.org

and provides a connection between controlled vocabularies (Courtot et al., 2011) and formats, tools, and databases. Most modeling standards provide mechanisms to attach MIRIAM annotations to their components. These annotations are structured based on a subject-predicate-object scheme. Here, the subject is the identifier of the model element. The predicate is one of several predefined qualifiers, e.g., hasPart or is. The object should be a web resources pointing to an identifiers.org address (Juty et al., 2012, 2013), for instance http://identifiers.org/kegg.compound/C00001 for water. This Uniform Resource Identifier (URI) is therefore composed of the prefix identifiers.org/, the definition of the data collection (in this example kegg.compound) followed by the delimiter and finally the record identifier (here C00001). Using such an identifiers.org address instead of directly pointing to an entry in ChEBI (Brooksbank et al., 2013; Hastings et al., 2013), MetaCyc (Caspi et al., 2014), KEGG (Kanehisa and Goto, 2000), or any other of the more than 30 currently supported data collections has several advantages. Should the original resource location or address schema change, the identifiers.org site will point to the new location. identifiers.org also measures the uptime of mirror servers for identical records and preferably directs to the most reliable mirror.

2.1.2. Minimum information about a simulation experiment

The minimum information about a simulation experiment (MIASE) project (Waltemath et al., 2011a) aims to unambiguously define how to reproduce the results of a model simulation. For stochastic models, the results should be within an acceptable small range from the original results, and for deterministic models, the results should be identical. This requirements checklist also supports the review process of scientific publications. Relevant ontologies (Courtot et al., 2011) for MIASE are the kinetic simulation algorithm ontology (KiSAO) that defines the method to use, the terminology for the description of dynamics (TEDDY), and the mathematical modeling ontology (MAMO).

2.1.3. Ontologies

2.1.3.1. Kinetic simulation algorithm ontology. The KiSAO gathers computational methods that can be used to simulate a model in a certain way (Courtot et al., 2011). It contains, for instance, definitions of several differential equation solvers for numerical calculations. Organizing these algorithms in a hierarchical structure allows tools to automatically select the most similar solver within their collection of implemented methods.

2.1.3.2. SBO. The Systems Biology Ontology is a collection of terms that describe the structure of a model, its components, modeling frameworks, and processes (Courtot et al., 2011). By using terms from this ontology, the semantics of individual parts of a model can be made explicit. This is often of particular importance if elements can participate in processes where they can have multiple roles, such as catalysts or inhibitors.

2.1.3.3. Mathematical modeling ontology. The recently developed ontology (MAMO, see http://bioportal.bioontology.org/ontologies/MAMO) has complemented and refined the *modeling framework* branch of SBO. Both ontologies are intended to cross link each other. While SBO mainly focuses on the entities and parameters in the model and describes the relationships among them, MAMO has been developed in order to precisely define and categorize *types* of mathematical models (e.g., *ODE*) and their characteristics (e.g., *discrete* vs. *continuous*) as well as types of readouts (such as *time-course analysis*) and variables (such as *dependent variable*).

2.1.3.4. Terminology for the description of dynamics. The TEDDY defines a formal way to specify how the numerical results of a dynamic system behave when a simulation experiment is conducted (Knüpfer et al., 2006; Courtot et al., 2011). In this way, a machine-readable representation of such a description can be automatically generated upon simulation and be stored along with the model. When querying a database of numeric results, this terminology can help to find models with a desired behavior, such as ongoing oscillations.

2.2. MODELING FORMATS

Reconstructing computational models based on a textual description in a publication can be difficult, because required information, such as a clear definition of the units of all components, can be lacking, the language might be imprecise or ambiguous, or a combined explanation of simulation procedure and actual model hamper the implementation of the model (Cooling, 2010; Dräger et al., 2010). In cases, where models are distributed in form of source code implemented for a specific run-time environment or programing language, executing these programs can be a challenge because of diverse dependencies to operating systems or required third-party libraries (Lloyd et al., 2004). In this section, we will discuss several formats that encode systems biological models in different ways with the aim to overcome this problem.

2.2.1. Systems biology mark-up language

The Systems Biology Mark-up Language (Finney and Hucka, 2003; Hucka et al., 2003, 2004)[2] is a hierarchical XML-based format consisting of several lists of components, such as compartments (finite spaces), (reactive) species, parameters (constants or variables), reactions with kinetic laws, user-defined functions and rules, events, units, and many more. SBML has been developed as a model exchange format that covers a wide range of modeling approaches used today (Hucka et al., 2004), including dynamic and steady-state metabolic networks as well as gene-regulatory and signaling networks (Lambeck et al., 2010; Vlaic et al., 2013). The term *reaction* should no longer be seen as a strict (bio-)chemical reaction. It is rather a process with inputs and outcomes. Specific annotation with SBO terms and MIRIAM identifiers clarify the purpose of all elements. The reactions implicitly define a differential equation system, whose explicit structure needs to be assembled at simulation time or prior to simulation. The rationale behind this design decision is that the same model can be interpreted in terms of a different modeling framework, such as stochastic simulation, etc.

The libraries libSBML (Bornstein et al., 2008) and JSBML (Dräger et al., 2011) facilitate the implementation of import and

[2] http://sbml.org

export functions of SBML models in customized software solutions. While libSBML provides bindings to a large variety of programing languages based on the wrapper generator SWIG (Beazley, 1996), the JSBML library has been specifically developed for the platform-independent Java™ language. Both libraries strive to attain a high degree of compatibility. Specific API libraries have also been implemented for working with SBML under MATLAB™ (Keating et al., 2006) and Mathematica™ (Shapiro et al., 2004, 2007).

It has been recognized that the interpretation and simulation of SBML models can be quite challenging and that different simulation environments can yield divergent results on identical input files (Bergmann and Sauro, 2008). For this reason, a comprehensive test suite of manually created SBML models has been established including reference results. This test suite can be used as a benchmark test case for simulation routines.

SBML handles the increasing diversification of modeling approaches and community requirements with the development of several specific and orthogonal packages, which can be used in addition or separately from the core format. The following extension packages have already been released: hierarchical model composition (comp) (Smith et al., 2013b), flux balance constraints (fbc) (Orth et al., 2010; Bergmann and Olivier, 2013), three-dimensional arrangement of elements in diagrams (layout) (Gauges et al., 2006), and qualitative relationships (qual) (Chaouiya et al., 2013). Draft specifications are available for the following extensions: arrays, sampling of values from statistical distributions (distrib), dynamic creation and destruction of structures during a simulation (dyn), grouping of elements (groups), entity pools with multiple states and complex composition of species (multi), drawing graphical representations of a model (render), indication of those model elements that are changed by packages (req), and spatial processes and geometries (spatial). For an up-to-date list and more detailed explanation of available extension packages, see http://sbml.org/Community/Wiki.

2.2.2. CellML

The XML-based model storage and exchange format CellML[3] has been developed for the IUPS Physiome project with the aim to facilitate reuse of models or their components in a software-independent manner (Lloyd et al., 2004; Cooling, 2010). CellML eases the creation of new models based on parts of existing models and hence accelerates the cumbersome model building process (Cooling et al., 2010). CellML models contain structural information about the organization of the model (components, connections, and units), mathematical equations (arbitrary MathML) to quantitatively describe biological processes, and metadata that link model components to online resources. An important design feature of CellML allows components and parameters to be shared across models via import statements and well-defined interfaces. This also allows users to structure their models into multiple files, similar as can be done with HTML pages, and increases reusability of individual black-box models, but also requires a strict decoupling of components. CellML uses RDF tags for semantic annotations and allows for hierarchical groupings of components. A set of software tools is available to edit CellML models, including an API implementation (Miller et al., 2010) or the graphical modeling environments OpenCell (Lloyd, 2013) and OpenCOR (Nickerson et al., 2013). CellML can be inter-converted from and to SBML and to the scripting language Antimony (Schilstra et al., 2006; Smith et al., 2013a). The rates of change of all components are explicit in CellML. When adding components or connections to a model, these rates of change would need to be updated. With the help of interfaces modelers can avoid this cumbersome update process (Cooling, 2010).

2.2.3. FieldML

FieldML[4] is an XML-based model interchange standard, which has been developed with a focus on the euHeart and Physiome projects and is currently available in version 0.5 (Britten et al., 2013). The main purpose of the format is to encode geometric models in explicit or implicit mathematical form with respect to biological and medical phenomena with spatial-temporal variation, such as the simulation of power fields and gradients. FieldML focuses on fields over multiple discrete indices and multivariate fields with discrete or continuous variables as well as interpolation functions. With these approaches, it is possible to model muscle contraction as part of cardiac mechanics, blood flows, and other multi-scale processes. Other applications include the modeling of patient-specific clinical images with the help of specific annotations and fitting of models to fields. Similar approaches are also planned for the spatial extension for SBML (Schaff et al., 2013). A powerful C++ API with wrappers for Java, Fortran, and Python as well as a software plug-in for the physiome model repository (PMR) support FieldML and provide several high-level functions for model building and simulation (Yu et al., 2011). Version 0.5 already includes model composition over multiple files and data sources.

2.2.4. BioPAX

The motivation for the creation of the BioPAX[5] format (Biological Pathway Exchange) was the aim to unify the various co-existing pathway encoding formats of numerous online databases (Demir et al., 2010). This format is intended to facilitate the communication between diverse software systems and also serves as a common knowledge representation of pathways. With BioPAX the structure of metabolic, signaling, and gene-regulatory pathways can be encoded, including relationships between elements (such as genes or molecules) as well as diverse states (such as post-translational modifications). A growing number of pathway databases and software tools provide BioPAX files as import or export formats (Shannon et al., 2003; Funahashi et al., 2008; Demir et al., 2010; Kelder et al., 2011) and BioPAX is useful to integrate information from heterogeneous sources, to support visualization, and analyses. The definition of BioPAX is the result of a continuous community effort. The BioPAX language is organized in levels that increasingly add features to the language definition. BioPAX is based on OWL and it is implemented as an ontology. An online

[3] http://www.cellml.org

[4] http://physiomeproject.org/software/fieldml

[5] http://www.biopax.org

validator can be used to check the correctness of BioPAX files. All elements within a BioPAX file can be annotated using controlled vocabularies and MIRIAM (Laible and Le Novère, 2007; Juty et al., 2012). For writing, reading, manipulating, and analyzing the API library Paxtools (Demir et al., 2013)[6] has been created and is freely available. Quantitative relationships and temporal sequences of events do not belong to the objectives of BioPAX. However, since it is also possible to encode qualitative relationships in SBML (Chaouiya et al., 2013), BioPAX can be converted to SBML without loss of information (Büchel et al., 2012).

2.2.5. *NeuroML*

The object-oriented mark-up language NeuroML (Gleeson et al., 2010)[7] has been developed as a standard to specifically encode, share, and store computational models of information transfer in neurosciences (Goddard et al., 2001). The aim of the language is to cover diverse structural levels beginning at individual neuron cell membranes and ranging to entire neural networks. This XML-based language encodes biophysically detailed neuronal and network models including ion channels, synapses, and the anatomical connectivity of neurons and how these elements underlie the complex electrical behavior of the brain (Gleeson et al., 2010). Therefore, from the very beginning, modularity, portability, and clarity were the main language requirements (Goddard et al., 2001). Supporting high-performance simulations and creating software frameworks for neuroinformatics are the aims of the language (Beeman, 2013). To this end, NeuroML 2 has been built on the Low Entropy Model Specification (LEMS) language (Cannon et al., 2014), which hierarchically defines structure and dynamics of a large variety of biological models. For parsing, writing, and manipulating NeuroML and LEMS files, the Python APIs libNeuroML and PyLEMS as well as the Java™ APIs jNeuroML and jLEMS are available (Vella et al., 2014). The original idea to link sub-modules of processes in NeuroML to models encoded in SBML or CellML (Gleeson et al., 2010) has since been further elaborated. The LEMS libraries allow users to import SBML models and can also export SED-ML (Waltemath et al., 2011b) files for reproducible simulation experiments. The main repository for NeuroML is Open-Source Brain (Gleeson et al., 2013).

2.2.6. *ISML and PHML*

The XML-based language ISML (*insilicoML*) allows users to describe biophysiological models that cross multiple scales and levels. This format is fully compatible to CellML 1.0, but incorporates a specific ontology of physiological functions (Asai et al., 2008). A large collection of models in ISML can be obtained from an online database at http://www.physiome.jp. The physiological hierarchy mark-up language (PHML) has been designed as the successor of ISML (Asai et al., 2013). PHML defines each biological or biophysical element as a module, which can be encapsulated and linked through ports. This concept hierarchically structures the language. Furthermore, PHML can integrate SBML models as sub-cellular phenomena (Asai et al., 2012).

2.2.7. *PharmML*

The Pharmacometrics Mark-up Language PharmML (Moodie et al., 2013)[8] belongs to the most recent languages in the family of XML-based standards for biomedical computation and is currently under development. The purpose of this new language is to exchange and store pharmacometric models, which includes studies, trials, simulations, estimation, and exploration. It will support metadata, non-linear mixed effects models, serve as an encoding platform for new approaches and elements, as well as support model-based analysis. The developers want to ensure backwards compatibility with existing relevant standards in order to use existing software tools. Use-case scenarios are, for instance, the kinetics of tumor growth, observation models, or trial design for treatment-dosing-related data.

2.2.8. *Synthetic biology open language*

The Synthetic Biology Open Language[9] also belongs to the latest modeling standards (Galdzicki et al., 2014). This RDF-based format has been designed in a community process in order to facilitate the creation of synthetic biology components by providing an exchange format for software tools. As a specialty, SBOL comes with a specific graphical representation for promoters, their regulators, and many additional genetic structures (see **Figure 2**).

2.3. STANDARDS FOR MODEL SIMULATION PROCEDURES

Defining the structure of a model does not give any information about reproducible simulation experiments. In order to perform the identical simulation of the model as described in a corresponding research article, the exact name of the numerical solving algorithm, step size, error tolerance, etc. must be precisely defined. The purpose of the Simulation Experiment Description Mark-up Language (SED-ML)[10] is to provide a standardized, machine-readable, platform-independent data format for this purpose (Waltemath et al., 2011b). SED-ML follows the MIASE guidelines (Waltemath et al., 2011a) and hence enables users to attach both a model as well as the description of its intended use to a publication, which could also simplify review processes. It therefore contributes to the reproducibility aspect in science, where only stochastic approaches might diverge within a small range from published data. The XML-based language SED-ML is organized in levels and can describe

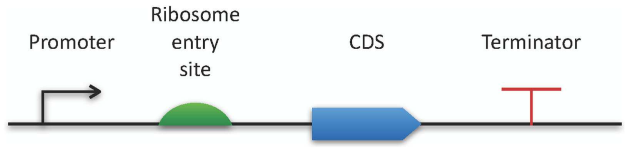


FIGURE 2 | SBOL visual. The horizontal bar represents a DNA molecule to which various features can be visually attached. Here, a few examples are applied for demonstration purposes. A full specification and an exhaustive list of all available symbols can be found online at http://www.sbolstandard.org/visual.

[6] http://www.biopax.org/paxtools/

[7] http://www.neuroml.org

[8] http://www.ddmore.eu

[9] http://www.sbolstandard.org

[10] http://sed-ml.org

multiple simulation experiments within the same file. Language components can be annotated using MIRIAM resources (Laible and Le Novère, 2007). A key idea of SED-ML is not to distribute concrete implementations of simulation procedures, but rather to use ontologies such as KiSAO (Courtot et al., 2011) to refer to the method and its settings. Since this ontology has a hierarchical structure, it is possible to apply related simulation algorithms in case a required method is not implemented in a certain software tool. Structural model changes prior to simulation and post-processing steps of the results (such as converting between amounts and concentration units) as well as the presentation of the output can also be defined (Waltemath et al., 2013). The model can, in principle, be encoded in an arbitrary standardized format and addressed through URI links. SED-ML does not provide an encoding of the simulation results itself, but can be used in combination with numerical mark-up language (NuML) or SBRGML (Dada et al., 2010). An extension to SED-ML has been proposed in order to also support sampling sensitivity analysis simulation experiments (Miller et al., 2012). Some simulation environments have already adopted this young format (Olivier et al., 2005; Myers et al., 2009; Kolpakov et al., 2011; Keller et al., 2013). A workflow editor (SED-ED), API libraries (libSedML, jlibSEDML), and a simplified scripting language (Antimony) are also available (Smith et al., 2009; Adams, 2012).

2.4. GRAPHICAL MODEL REPRESENTATION FORMATS

The visual representation of biochemical pathways has a long tradition. Displays of biological circuit diagrams and reaction pathways can be found in numerous textbooks and a plethora of publications. Databases such as KEGG (Kanehisa and Goto, 2000) or MetaCyc (Caspi et al., 2014) take this up and provide displays of biological networks in their specific layout and style, which follows many traditional aspects. In order to display and draw similar maps, several programs have been developed, for instance, CellDesigner (Funahashi et al., 2008), JDesigner (Sauro et al., 2003), TinkerCell (Chandran et al., 2009), VCell (Resasco et al., 2012), or Cytoscape (Shannon et al., 2003) with its diverse plug-ins (König et al., 2012; Gonçalves et al., 2013). We now discuss recommendations for the display of pathways and standardized data formats for exchanging these maps.

2.4.1. SBGN and SBGN-ML

The myriad of graphical notations that are being used can lead to confusion or ambiguity. The development of a unified and standardized notation has thus become necessary (Le Novère et al., 2009). The Systems Biology Graphical Notation[11] effort aims to make the display of biological networks exchangeable between software tools and at the same time to clearly define the meaning of specific nodes and arcs in such networks in order to ease their interpretation and automated processing. Therefore, the number of graphical symbols is intentionally limited in order to keep the learning curve flat and to create a visually, syntactically, and semantically consistent schema, which is modular in size and complexity (Le Novère et al., 2009). The SBGN neither defines layout (placement and adjustment) nor style (such as line thickness or color) of objects. In order to represent the current needs for such a display, it is organized in levels, so that in the future new versions can be proposed. The specifications of the SBGN are organized in three different languages, each of which has been designed for certain use-case scenarios and has inherent strengths and weaknesses. (i) In process-description diagrams (Kitano et al., 2005; Funahashi et al., 2008), the level of detail is very high and these maps show sequences of processes, which also involve temporal causality (see **Figure 3A**). These maps are well suited for metabolic pathways, but not for the consistent display of the combinatorial complexity of several proteins with many phosphorylation states (van Iersel et al., 2012). (ii) Activity flow charts (van Iersel et al., 2012) are much more abstract and neglect many molecular mechanisms. By design, these maps introduce a certain ambiguity and can hence be used to describe effects whose precise underlying mechanisms are either not know or not relevant (see **Figure 3B**). In this type of diagram, stimulation and inhibition, effects of perturbation, and the activity of components can be displayed. Activity flow charts are thus suitable for the display of causality chains (van Iersel et al., 2012). (iii) The entity-relationship diagrams (Kohn et al., 2006) are particularly useful when the temporal sequence of events does not play the main role, but precise molecular interactions are to be displayed (see **Figure 3C**). These maps are more concise than process-diagrams for protein modifications and interactions, but less capable of representing reactions (van Iersel et al., 2012).

In order to specifically store and exchange SBGN maps in XML files, the Mark-up Language SBGN-ML has been developed (van Iersel et al., 2012). The main requirement for this format is its simplicity, i.e., it should be easy to draw and to interpret. Most significantly, SBGN-ML is not tied to any of the network representation standards. While, this format does not include rendering information, it has been proposed to incorporate a rendering extension, similar as can be done with SBML files. In contrast to the SBML layout extension, this format is focused on the concepts of SBGN only and can be validated against the SBGN specifications. The API library libSBGN[12] facilitates the import and export of SBGN-ML files. The code of libSBGN has been automatically created from an XML Schema Definition file (XSD), which significantly reduces the implementation effort, makes native language implementations in C++ and Java™ possible, and can be used for Schematron validation. A growing number of libSBGN-based software tools support the SBGN-ML format, such as the VANTED (Junker et al., 2006) plug-in SBGN-Ed (Czauderna et al., 2010), the Cytoscape (Shannon et al., 2003) plug-in CySBGN (Gonçalves et al., 2013), the online tool BioGrapher (Krause et al., 2013), the model generator KEGGtranslator (Wrzodek et al., 2013), or the visual editor CellDesigner (Funahashi et al., 2008).

2.4.2. Visualization of CellML

For CellML, a specialized interactive framework has been developed for the display of models (Wimalaratne et al., 2009). This

[11] http://sbgn.org

[12] http://libsbgn.sourceforge.net

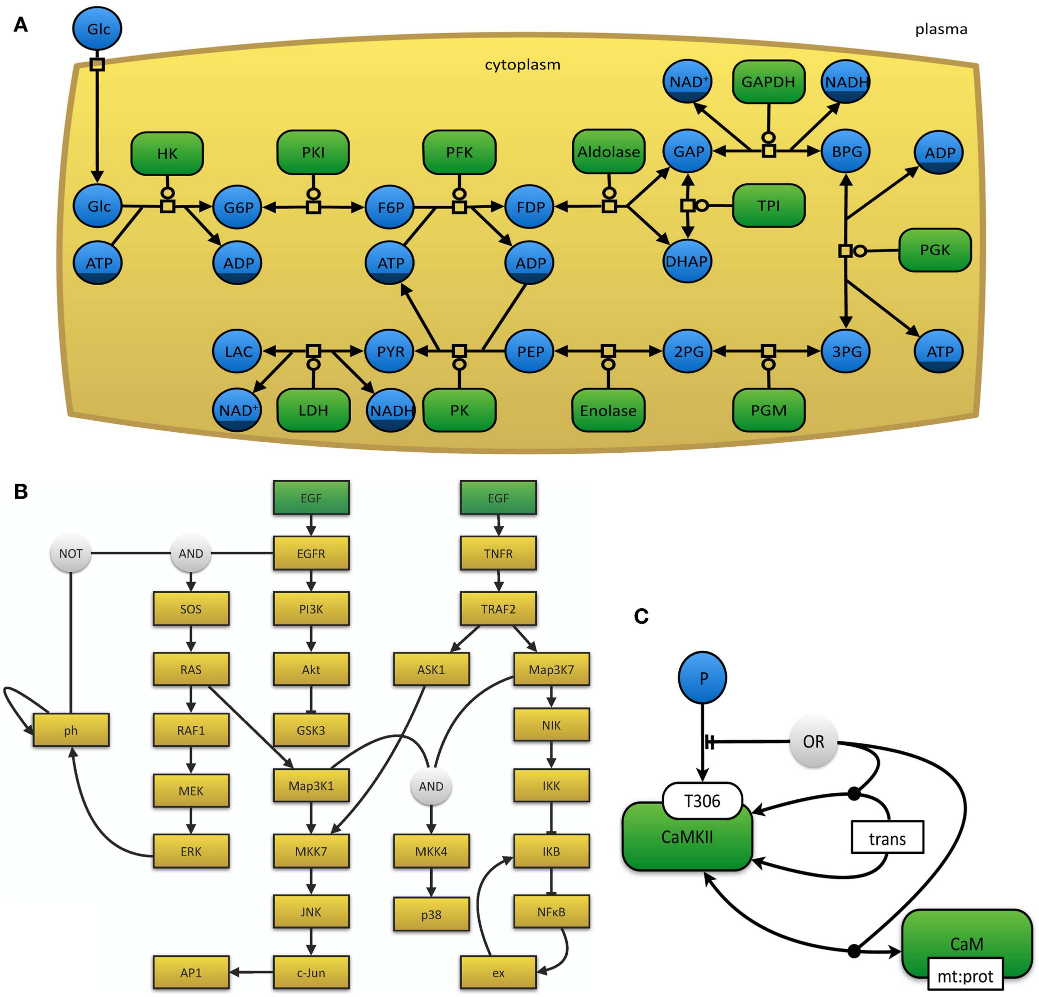


FIGURE 3 | (A) The glycolysis in human erythrocytes, simplified from Dräger (2011). This example network depicts the reaction steps from extracellular glucose to intracellular lactose as a chain of enzyme-catalyzed reactions in SBGN PD notation. Metabolites that occur multiple times in the map, such as ATP or NAD$^+$, have darker clone markers on the bottom. Simple molecules are displayed as circles, whereas, macromolecules appear as rounded rectangles. Reactions are indicated as square process nodes. **(B)** This activity flow diagram displays the interaction of two mammalian signaling pathways that are stimulated by epidermal growth factor (EGF) and tumor necrosis factor alpha (TNFα) and their influence on the nuclear factor κ-light-chain-enhancer of activated B cells (NFκB) and mitogen-activated protein kinases (MAPK) cascades. Adapted from Chaouiya et al. (2013) and generated with the program CellNOpt (Terfve et al., 2012). Here, external stimuli are colored in green. **(C)** This figure displays an example for an entity-relationship diagram, in which Ca^{2+}/calmodulin-dependent protein kinase II (CaMKII) is precluded if it dimerizes or if it binds to the protein calmodulin. Adapted from Le Novere et al. (2011).

framework can either depict the physical model, i.e., the actual components of the CellML format, or the biological interpretation. CellML hence provides its own two-dimensional visual language for both concepts, which can be used in programs to link between the display and the underlying data structure, and also for dynamic image manipulation. For both kinds of displays, a small set of distinct glyphs are defined: entities, processes, and roles. While the physical display tends to be very complex, the biological view is much more straightforward. The developers of the CellML visualization scheme interact with the SBGN team (Wimalaratne et al., 2009). On the longer term, it is intended to combine ideas from SBGN (Le Novère et al., 2009) and the CellML display. Currently, not all concepts of the CellML display can be expressed in SBGN (Wimalaratne et al., 2009).

2.4.3. Layouts in SBML

Layouts can directly be stored in SBML models with the help of the layout extension (Gauges et al., 2006). With this extension, it is possible to attach information about position and size of objects, such as reactive species, compartments, or reaction arcs. Text labels can also be placed. The SBML layout package is based on boundary boxes and defines neither shapes nor colors of objects, but it can be further extended with additional rendering information (Deckard et al., 2006; Shen et al., 2010). Tools such as SBML2LaTeX or SBML2Ti*k*Z (Dräger et al., 2009b; Shen et al., 2010) can interpret layouts stored in this extended SBML to be consistent with SBGN process-diagram maps. In general, these two SBML extensions allow users to store arbitrary forms of network representations. Programs such as KEGGtranslator

(Wrzodek et al., 2011, 2013) use the layout extension to preserve initial layouts from the KEGG database in SBML files. In combination with the SBML extension for qualitative models (Chaouiya et al., 2013), it is also possible to create activity flow networks. In contrast to the SBML layout extension, no standardized way has been proposed to directly store SBGN-ML layouts inside of SBML files. However, the recent COMBINE format (Bergmann et al., 2014) allows users to store files of diverse forms all together within one archive file (see Section 2.6).

2.5. REPRESENTATION OF NUMERICAL RESULTS

In order to store the results of numerical simulation, specific file formats have been proposed. The Systems Biology Result Mark-up Language (SBRML, Dada et al., 2010) has been succeeded by the NuML[13]. This new format has been developed as a standardized exchange and archiving format for the results of numerical methods. This new language has been designed as a format that is usable in various disciplines besides systems biology. The C++ library libNUML can be used for parsing, manipulating, and writing the information of NuML data structures.

2.6. COMBINE FORMAT

The COMBINE format aims to distribute diverse modeling, documentation, and data files together within one single Open Modeling Exchange format (OMEX) file (Bergmann et al., 2014). The format is basically a ZIP archive, i.e., a compressed datatype, which contains an XML-based manifest file and an optional metadata file in RDF format. While the structure of the manifest file is well-defined, there are only recommendations for the metadata file. If present, metadata should at least include information about the author of the OMEX file in form of a vCard and follow the structure proposed by the Dublin Core Metadata Initiative. The manifest file contains structured links to all included files together with a definition URI that describes the filetype. Thus, diverse types of files can be included, even publications, plots, models, graph definitions, etc. Just for the sake of significant data compression, it is already recommended to store models inside of OMEX files (file extension *.omex). Even though the COMBINE archive format belongs to the most recent datatypes of the systems biology community, it is already supported by a number of tools and also the library libCombineArchive for dealing with it (Java™ and C#).

2.7. ONLINE MODEL REPOSITORIES AND DATABASES

One important aspect of model exchange and reusability is the availability and distribution of models that have already been published or that are currently under review. Since a growing number of journals require the online availability of models along with a publication, it is important to be familiar with a number of online resources that are now available. In this section, we will discuss the different aims and features of selected online model repositories, which are summarized in **Table 2**.

2.7.1. BiGG

An important resource for Biochemically, Genetically, and Genomically structured genome-scale metabolic network reconstruction is the BiGG database (Schellenberger et al., 2010). The main focus of this knowledge-base is to facilitate the bottom-up genome-scale reconstruction of metabolic networks. Inclusion of every known reaction of an entire organism constitutes the ultimate goal of BiGG. To this end, it integrates published genome-scale metabolic networks into one resource and applies a standard nomenclature for all of their components. Among these networks are several important model organisms, such as *E. coli* and *H. sapiens*, as well as further main branches of life (Duarte et al., 2007; Feist et al., 2007; Thiele et al., 2013). All models are manually curated and all reactions are atom-balanced. These networks also include gene–protein associations, which can be used to relate the activity of genes via Boolean logic to reactions and hence to perform knock-out or knock-down experiments *in silico*. BiGG offers various options to search, browse, and display networks. Manually curated maps can be downloaded in SVG format for a multitude of pathways. There are often several such maps available for one organism. Various build-in functions (such as decompartmentalization, orphan detection, gap filling, etc.) support the modeling process. With its SBML export function, it provides the basis for further steps in the modeling pipeline, particularly constraint-based analyses by the COBRA platform (Becker et al., 2007; Ebrahim et al., 2013). As the first database specific to constraint-based models, it precedes the SBML extension for fbc,

Table 2 | Relevant online databases.

Database	URL	Provides	Comments
BiGG	http://bigg.ucsd.edu	SBML	COBRA models
BioModels	http://www.ebi.ac.uk/biomodels-main/	CellML, SBML, PDF, VCML, and other formats	Main repository for SBML models
JWS	http://jjj.biochem.sun.ac.za/	JWS format, SBML	Online simulation facility
ModelDB	http://senselab.med.yale.edu/modeldb/	Various kinds of model data files	Focus on neuroscience
Open-source brain	http://www.opensourcebrain.org	NeuroML and PyNN	Interactive model development repository
PMR2	http://models.cellml.org	CellML	Project management platform with connection to JWS
SEEK	http://www.sysmo-db.org	Models in diverse formats, publications, and presentations	Focus on collaboration, connection to JWS
WikiPathways	http://www.wikipathways.org	BioPAX, PathVisio, and image formats	Interactive web 2.0 tool for biochemical pathways

[13] http://code.google.com/p/numl/

but provides COBRA-specific model extensions that can be easily converted (Bornstein et al., 2008).

2.7.2. *BioModels database*

BioModels database (Chelliah et al., 2013) is an open-source project, whose license model allows free commercial and academic use. Individual authors can submit their models to this database. A team of curators further improves the models, for instance by making the annotations in the model consistent with respect to MIRIAM guidelines (Juty et al., 2012). Large parts of the database content have been imported from collaborative repositories, such as the CellML model repository (Yu et al., 2011). The web interface of BioModels database provides a large variety of services based on embedded tools, e.g., for the simulation or graphical display of models. The main format of BioModels database is SBML, but models can be downloaded in a wide variety of formats, most of which have been automatically converted from the SBML files. It is also possible to obtain an exhaustive model report about each model (Dräger et al., 2009b) that describes the details of each model component in a human-readable way. Since the database was launched in 2005, it has been observed that not only are the number of models significantly increasing, but also their complexity. It now contains a large number of models, each describing the same biological process, but with higher levels of detail. With the growing size of the database the search for a model of interest has become a problem by itself (Schulz et al., 2011). With the help of metadata stored along with each model and the actual content of the models, sophisticated ranking procedures have been designed based on information theory aiming to retrieve models from the database for a given query (Henkel et al., 2010). The metadata include the submission and modification data, the authors of the model, and references. The user can browse through the models based on several characteristics, including the model name, publication identifier, or a GO-based (Ashburner et al., 2000) classification. Besides the curation of models, the main purpose of this repository includes the reproduction of model simulation results as given by the original publication (Waltemath et al., 2013).

2.7.3. *CellML physiome model repository 2*

The CellML physiome model repository 2 (PMR2) is the most important resource for CellML models at different states of their curation (Yu et al., 2011). It uses a Plone-based model management system that is organized in workspaces. This allows its users to collaboratively develop models based on a version-control system and also facilitates the modular development of models. The models stored in this database cover a large variety of processes, including signal transduction and metabolic pathways, electrophysiological and cell cycle models, immunological models, and models describing muscle contraction or mechanical phenomena. The idea of collaborative model development brings with it one important feature: PMR2 keeps track of a detailed version history of all models. Plug-ins to the system facilitate the presentation of models in various ways and also enable the import and export of diverse modeling formats, including SBML or FieldML besides the native database format CellML. In addition, the plug-in technique makes the database extendable. A search function returns models of all curation states. The main focus of this database is to provide a version-controlled repository for the collaborative model development and presentation of model information, here called *exposures*.

2.7.4. *JWS online model repository*

Another popular model resource is the JWS Online Model Repository (Snoep and Olivier, 2003). When JWS was launched as the first central model database in 2003 the standards SBML and CellML were still in their early development and not as well established. The repository itself is tightly related to the JWS online simulator (Olivier and Snoep, 2004), a particularly useful resource for educational purposes. Since then, the database has been continuously extended. Its native data format is SBML. Models can be queried based on a list of predefined characteristics (Waltemath et al., 2013), including metadata such as author, publication, organism, or model type as well as a list of categories (for instance, cell cycle or metabolism). The purpose of JWS is to provide a user-friendly online repository of kinetic models of biological systems in combination with an application that facilitates the simulation of these models. The aim of this infrastructure is to ease the review process of papers describing these kinds of models. As a result of its integration into the SEEK platform (Wolstencroft et al., 2011) a large number of collaborative projects use JWS as their default modeling platform.

2.7.5. *SEEK platform*

The open-source SEEK platform benefits from the ability to offer JWS as its integrated simulation tool to its users. The SEEK platform goes beyond just being a model database. This web-based tool has been designed as a pragmatic data management solution for the exchange of very diverse kinds of data relevant for research in systems biology. Besides mathematical models, it also covers the exchange of experimental data, scientific protocols, and personal information about members of large research consortia (Wolstencroft et al., 2011). It allows its users to record the outcomes of experiments. One of its most important features is the ability to link between data, models, and publications, as well as to tag all uploaded items. This platform has originally been developed for the European SysMO consortium (Booth, 2007), and is also used in several other National and European projects, such as the German Virtual Liver Network (Holzhütter et al., 2012). The preferred modeling data format of SEEK is SBML with MIRIAM annotations.

2.7.6. *ModelDB*

ModelDB (Migliore et al., 2003; Hines et al., 2004) belongs to the seven databases of SenseLab (NeuronDB, CellPropDB, ModelDB, Olfactory Receptor Database, OdorDB, OdorMapDB, and BrainPharm). SenseLab aims to provide a neural, genomics/genetics, proteomics, and imaging information resource for the neuroscience community and the interested public (Crasto et al., 2007). The database does not explicitly require a standard data format. Instead, authors are welcome to upload their models in arbitrary formats. As a result, the database is very flexible, but model reuse can take extra time to convert the desired model in a format for a particular execution environment (Waltemath et al., 2013).

2.7.7. Open-source brain

Inspired by the open-source movement, the collaboration-oriented open-source brain (OSB) repository has been established (Gleeson et al., 2013). All models in this repository can be commented, debugged, and extended by registered users. This platform therefore complements repositories, such as ModelDB (Hines et al., 2004), which focus on distributing published models, with the aim to drive the advance of models at all stages of its development. An integrated WebGL-based 3D explorer allows users to view cells and networks in NeuroML 2 format within their browser. OSB is well integrated and links to ongoing research projects such as OpenWorm[14].

2.7.8. WikiPathways

The WikiPathways project (Kelder et al., 2011) provides a Web 2.0 wiki-based platform for the online curation of biological pathways. The idea for this platform is that manually curated pathways are of higher quality than automatically created ones. Motivating the scientific community to share knowledge would thus increase the quality of available pathway information. To this end, WikiPathways provides an interactive zoom-able pathway viewer that comes with a pathway diagram description, hyper-links, and detailed information as well as literature references. Users can also annotate the pathways with ontology terms. It is possible to submit private pathway information that is shared later with the public, for instance, as part of the review process, or if current knowledge about certain processes is limited. As a major feature, WikiPathways provides stable hyper-links to all pathways, which is useful in order to use the platform as a reference. Its content can be downloaded in many export formats under the terms of the Creative Commons license. The BioPAX standard (Demir et al., 2010) is thereby its most important format. Internally, it uses GPML, an XML standard that is compatible with many modeling tools, including Cytoscape (Shannon et al., 2003).

3. RESULTS

3.1. INTEROPERABILITY OF STANDARDS

3.1.1. Path2Models

An important driving force for improved interoperability and exchange of diverse data formats and standards was the community project path2models (Büchel et al., 2013a). The aim of this project was to automatically create draft models of biological processes based on the knowledge stored in the databases KEGG (Kanehisa and Goto, 2000), MetaCyc (Caspi et al., 2014), SABIO-RK (Wittig et al., 2014), and PID (Schaefer et al., 2009). The extraction of information from these databases required the development of new algorithms in order to capture a large variety of special cases (Wrzodek et al., 2011, 2013; Büchel et al., 2012) due to the different scope of the source databases. In order to also encode qualitative networks in SBML, the standard needed to be extended (Chaouiya et al., 2013). The draft SBML models had to be quality controlled and enriched with further kinetic information for reactions for which the SABIO-RK database did not yet provide experimentally determined rate laws (Dräger et al., 2008, 2010). Drafts of whole organism models were created by combining individual organism-specific pathway models (Swainston et al., 2011).

The main purposes of the KEGG databases are to provide a comprehensive, textbook-like educational view on the knowledge about a large variety of biological pathways. For modeling purposes, however, the information needs to be presented in a different way (Wrzodek et al., 2013). Reactions cannot be lumped together for the purpose of a better visual presentation, but have to be made explicit. The model must be as specific as possible, i.e., organism-specific variations must be reflected in pathways.

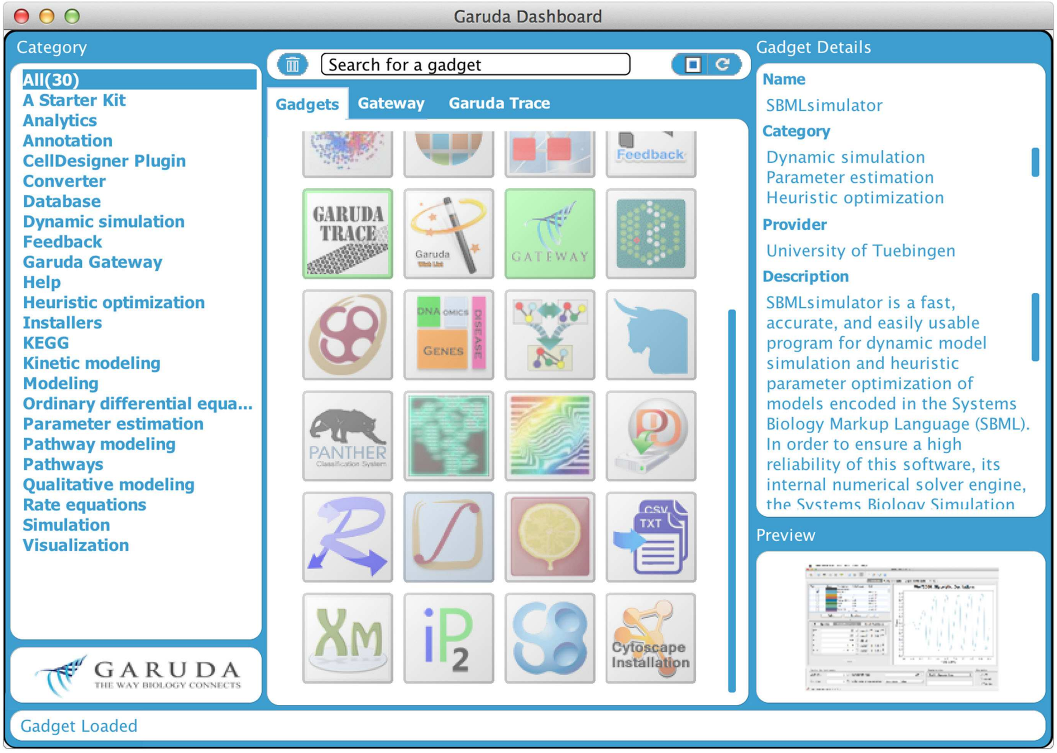
New algorithms also needed to be proposed in order to generate SBGN-ML files directly from KEGG (Czauderna et al., 2013). On the one hand, the manually created pathway maps in KEGG can be much better comprehended by human beholders than automatic layouts. However, in order to obtain an unambiguous representation of knowledge, the initial KEGG layout needs to be modified and subject to several constraints with respect to the esthetics of the result.

Such a large-scale endeavor, which resulted in more than 140,000 pathway maps that are all available from BioModels Database (Chelliah et al., 2013), was only feasible with the help of automatic procedures. Overall, this effort can be seen as a showcase application, which demonstrated the usefulness of data standardization, source code exchange, and software development in a large collaborative community project.

3.1.2. Workbench and workflow approaches

Even though several data storage and exchange formats have been defined and software has been developed to import and export those formats, it is still difficult to work with a large number of different programs and in diverse environments. It can be of particular interest to process intermediate results from one program in another software package or to work with software on different computers with different operating systems. Furthermore, software is often written in diverse programing languages and compiled in diverse environments. Code reuse is still quite limited. All this can hamper building complex analysis pipelines. To address these problems, the systems biology workbench (SBW, Sauro et al., 2003) and the Garuda effort (Ghosh et al., 2011) have been launched. SBW is a software framework for communication between heterogeneous application components. It provides a broker to which each SBW-enabled software needs to register. This broker enables the software to be executed on different machines. Information can be sent from one program to the other through a specific protocol, which provides a fast binary encoded message system. SBW therefore allows programs to use each other's capabilities. In contrast, Garuda is similar to an "App Store" for systems biology (see http://www.garuda-alliance.org/). It provides a common platform, from which diverse applications (gadgets) can be launched (see **Figure 4**). Garuda gadgets can call each other and send their output the next gadget or receive input from other gadgets. A powerful workflow would be to create a model with KEGGtranslator (Wrzodek et al., 2011, 2013), which can forward its result to the rate law generator SBMLsqueezer (Dräger et al., 2008, 2010), which in turn launches SBMLsimulator (Keller et al., 2013)in order to run a simulation and parameter calibration on

[14] http://openworm.org



Name
SBMLsimulator

Category
Dynamic simulation
Parameter estimation
Heuristic optimization

Provider
University of Tuebingen

Description
SBMLsimulator is a fast, accurate, and easily usable program for dynamic model simulation and heuristic parameter optimization of models encoded in the Systems Biology Markup Language (SBML). In order to ensure a high reliability of this software, its internal numerical solver engine, the Systems Biology Simulation

FIGURE 4 | Garuda dashboard. This is the main screen of Garuda. The left column lists several categories that group individual gadgets. The icons in the center column allow users to launch applications with a double click. A detailed description of a gadget is displayed in the right column upon click on an icon.

the resulting model. Garuda provides a nice and easily understandable user interface, the dashboard, from which applications can be launched.

3.2. SOFTWARE SUPPORT

A large variety of software has been developed for many kinds of model building, analysis, drawing, simulation, and format interconversion. In this section, we will only discuss a small number of conceptual categories and particularly important tools. Several reviews specifically focus on available software (e.g., Dandekar et al., 2012; Hamilton and Reed, 2013; Fernández-Castané et al., 2014; Gostner et al., 2014; Koussa et al., 2014; Kramer et al., 2014). **Table 3** gives an overview of selected software. For an up-to-date list and comprehensive information, see, for instance, the dynamic software matrix at http://sysbioapps.dyndns.org/pivot-software-matrix.html.

3.2.1. Visualization and model building

Several tools provide interactive graph-based user interfaces and facilitate import or creation, manipulation, or export of complex pathway structures. Some programs can be extended via plug-ins, e.g., the Biological Network Analyzer BiNA (Gerasch et al., 2014), CellDesigner (Funahashi et al., 2008), or Cytoscape (Shannon et al., 2003). The flexible stand-alone application BiNA (Gerasch et al., 2014) is based on a hierarchical graph concept and provides highly configurable styles for the visualization of regulatory and metabolic network data as well as access to the BN++ pathway data warehouse (Küntzer et al., 2007). The web-modeling tool BioGrapher (Krause et al., 2013) is implemented with HTML5, CSS, and JavaScript and can be used to create SBGN maps. BioGrapher can import several standard formats, including SBML and SBGN-ML, and export SBGN maps in a JSON file format or as images. The VANTED plug-in SBGN-ED supports all three kinds of SBGN maps and is therefore useful for designing and modifying SBGN-ML files (Czauderna et al., 2010). The framework program Cytoscape supports creation, import, and export of SBML and SBGN through plug-ins (König et al., 2012; Gonçalves et al., 2013). The main purpose of the straightforward and user-friendly process-diagram editor CellDesigner is the creation, manipulation, and simulation of SBML models (Matsuoka et al., 2014) with export functions to BioPAX (Mi et al., 2011) and SBGN-ML (van Iersel et al., 2012). CellDesigner can be extended through plug-ins, such as the kinetic law generator SBMLsqueezer (Dräger et al., 2008, 2010). The draft model generator KEGGtranslator (Wrzodek et al., 2011, 2013) automatically downloads contents of the pathway database KEGG (Kanehisa and Goto, 2000) and converts the content to diverse output formats, including SBML with extensions for layout (Gauges et al., 2006) and qual (Chaouiya et al., 2013), SBGN-ML (van Iersel et al., 2012), BioPAX (Demir et al., 2010), and many more. TinkerCell (Chandran et al., 2009) has been developed as a computer-aided design (CAD) tool and provides visual representations for systems biology and synthetic biology. OpenCOR (open-source cross-platform) for working with CellML files can be used through command-line or graphical user interface

Table 3 | Selected relevant software for systems biology.

Program	Main features	Citation
BiNA	Visualization of regulatory and metabolic network data with configurable styles and hierarchical graph concepts; analysis of omics data; data warehouse; plug-in system architecture	Gerasch et al. (2014)
BioGrapher	Web-based tool for creation and editing of SBGN maps with automatic layout algorithms	Krause et al. (2013)
BioUML	Platform for network building, simulation, analysis with full implementation of SBML	Kolpakov et al. (2011)
CellNOpt	Logic-based program for creating and simulating models of signal transduction	Terfve et al. (2012)
Cytoscape	Plug-in-based open-source software platform for visualizing complex networks and their attributes	Shannon et al. (2003)
CellDesigner	Process-diagram editor for gene-regulatory and biochemical networks with plug-in architecture and integrated solvers	Funahashi et al. (2008)
COBRA, COBRApy	Implementations of FBA, gene deletions, flux variability analysis, sampling, and batch simulations for constraint-based models	Schellenberger et al. (2011), Ebrahim et al. (2013)
COPASI	Simulation and analysis of biochemical networks and their dynamics in stochastic and ODE frameworks with support for SBW, parameter estimation, visualization, and several export formats	Hoops et al. (2006)
FASIMU	Command-line based collection of common FBA algorithms for SBML and several kinds of constraints. Its linear programing solvers can be exchanged and numerous constraints be defined	Hoppe et al. (2011)
Flint	An efficient stand-alone solver for PHML and SBML models, which also provides a cloud service	Asai et al. (2013)
GINsim	Simulator for qualitative gene interaction networks with graph-drawing capability, interactive user interface, and support for SBML qual	Gonzalez Gonzalez et al. (2006)
GRN2SBML	Converts the output of network inference procedures to SBML including MIRIAM annotation; access to BioMart central portal; R-package	Vlaic et al. (2013)
iBioSim	Modeling, analysis, design of genetic circuits for systems, and synthetic biology; user-friendly editors for diverse formats; variety of ODE and stochastic simulators; and plotting functions	Myers et al. (2009)
JSim	Building and analysis of quantitative numeric models with focus on physiology and biomedicine; support for ODEs, PDEs, implicit equations, etc.	Butterworth et al. (2014)
libRoad-Runner	C++ library for efficient numerical simulation and analysis of SBML models that provides Python language-bindings, which are integrated into the tellurium environment	Sauro et al. (2013)
libSBMLSim	C-based ODE simulation library for SBML models with explicit and implicit methods, language-bindings, and command-line tool	Takizawa et al. (2013)
Mass-Toolbox	Mathematica framework for kinetic and constraint-based model building and simulation; focus on mass-action kinetics and elementary reaction systems; support for ODE/DAE (incl. delays and events)	Sonnenschein and Palsson (2013)
Module-Master	Identification of *cis*-regulatory modules (CRMs) in sets of co-expressed genes based on transcription factor binding information and multivariate functional relationships between regulators and target genes	Wrzodek et al. (2010)
MOOSE	Multi-scale object-oriented simulation environment for diverse biological systems with a Python scripting interface and support for SBML, NeuroML, GENESIS kkit, and cell.p formats	Dudani et al. (2013)
OpenCOR	Plug-in based cross-platform modeling environment for working with CellML files	Nickerson et al. (2013)
Physio-Designer	Platform for the creation and analysis of PHML models that also allows users to integrate SBML models. It uses Flint as its solver back-end through a cloud service	Asai et al. (2013)
PySCeS	Extendable Python toolbox for time-course simulation, steady-state and stability analysis, metabolic control analysis and many more, support for SBML fbc and SED-ML	Olivier et al. (2005)

(Continued)

Table 3 | Continued

Program	Main features	Citation
SOSlib	C programing library for symbolic and numerical analysis of chemical reaction network models encoded in SBML format	Hindmarsh et al. (2005), Machné et al. (2006)
SBML-simulator	Dynamic model simulation and heuristic parameter optimization of SBML models based on the systems biology simulation core library and EvA2	Kronfeld (2008), Keller et al. (2013)
SBML-squeezer	Context-sensitive generator for kinetic equations of biochemical and gene-regulatory networks with access to SABIO-RK	Dräger et al. (2008, 2010), Dräger (2011)
SBToolbox2	MATLAB™ toolbox with support for SBML, and a large variety of analysis and high-performance simulation functions as well as parameter estimation, sensitivity analyses	Schmidt and Jirstrand (2006), Schmidt (2007)
TinkerCell	Computer-aided design platform for synthetic biology with C and Python API	Chandran et al. (2009)
VANTED	Versatile plug-in based visualization and analysis platform for networks with support for SBGN-ML, sophisticated layout algorithms, and FBA	Junker et al. (2006)
VCell	Modeling and simulation (deterministic and stochastic) of physicochemical and electro-physiological processes with support for irregular spatial distribution of substances in arbitrary geometries	Moraru et al. (2008), Resasco et al. (2012)

(Nickerson et al., 2013). It supports various aspects of modeling, including editing, simulation, and analysis. As a plug-in based program, OpenCOR can be easily extended. One of its most recent plug-ins facilitates the annotation of CellML.

3.2.2. *Constraint-based modeling*

The most important toolboxes for Constraint-Based Reconstruction and Analysis (Bordbar et al., 2014) are the COBRA Toolbox for MATLAB (Schellenberger et al., 2011) and its Python implementation COBRApy (Ebrahim et al., 2013). These toolboxes provide state-of-the-art implementations of flux balance analysis methods, including gene deletions, flux variability analysis, sampling, and batch simulations. Both versions of COBRA incorporate tools to read-in and manipulate constraint-based models, which requires a specific extension of the SBML standard. The Mathematica-based Mass-Toolbox (Sonnenschein and Palsson, 2013)[15] is a complex framework for constraint-based model building and simulation, which can calculate steady-state solutions for complex enzyme reactions and even solve ODE and DAE systems with delays and events. Further important tools for FBA are FASIMU (Hoppe et al., 2011), the VANTED (Junker et al., 2006) plug-in FBA-SimVis (Grafahrend-Belau et al., 2009), and PySCeS (Olivier et al., 2005).

3.2.3. *Dynamic simulation*

The main focus of the Mass-Toolbox (Palsson, 2011; Sonnenschein and Palsson, 2013) is kinetic modeling with a focus on mass-action rate laws and elementary reaction systems. It supports a large variety of analysis methods and high-level plotting commands for phaseportraits, and many more.

The SBToolbox2 (see http://sbtoolbox2.org, Schmidt and Jirstrand, 2006; Schmidt, 2007) provides a powerful and extensible variety of simulation and analysis functions, which smoothly integrate into the MATLAB environment. SBToolbox2 supports SBML and parameter estimation with EvA2 (Kronfeld, 2008).

CellDesigner delivers several third-party tools for interactive model simulation SOSlib (Machné et al., 2006), the Simulation Core Library (Keller et al., 2013), or COPASI (Hoops et al., 2006).

The SBW-enabled complex pathway simulation program COPASI is primarily a stand-alone program, but provides API language-bindings for several programing languages. COPASI can read, write, and understand SBML, but has its own specific modeling language and supports several other export formats. It comprises methods for simulation and analysis of biochemical networks and their dynamics based on ODEs and stochastic systems. Parameter estimation and the visualization of data as well as animated pathways are among its strengths.

The tool SBMLsimulator combines the Simulation Core Library, a comprehensive Java™ API for solving SBML models (Keller et al., 2013) with the optimization framework EvA2 (Kronfeld, 2008) in a self-explanatory user interface and provides a complete implementation of the SBML standard in terms of an ODE framework.

The stand-alone desktop tool BioUML (Kolpakov et al., 2011) is among the few tools that provide a full implementation of the SBML standard in terms of ODE systems and also provides its functions as JavaScript API.

The stand-alone tool iBioSim (Myers et al., 2009) for modeling, analysis, and design of genetic circuits has been developed as an editor and simulator (ODE and stochastic) with applications in systems biology as well as synthetic biology. Besides SBML, it also understands Petri net (LPN) models and has import access to model databases. Experimental data can also be used to infer models in iBioSim.

SOSlib (Machné et al., 2006) is an ODE-based C-API library implementation of SBML that internally uses CVODE (Hindmarsh et al., 2005). The newer C-implementation libSBMLSim (Takizawa et al., 2013) supports even more recent versions of SBML, explicit and implicit integration methods, and bindings to several programing languages. Another alternative is libRoadRunner, a highly performant C++ library for the simulation of

[15] http://opencobra.github.io/MASS-Toolbox/

SBML models, which provides automatically generated language-bindings to Python (Sauro et al., 2013).

The Java-based tool JSim has been designed for building quantitative numeric models as well as the analysis of these models based on given experimental data (Butterworth et al., 2014). It supports ODEs and PDEs, discrete events, and implicit methods. JSim can import and export SBML and import CellML (Smith et al., 2013a).

The Virtual Cell suite VCell is a powerful simulation toolbox for complex biological phenomena (Moraru et al., 2008; Resasco et al., 2012). It includes sophisticated methods for: (i) molecular interactions and transport, (ii) various sub-cellular compartments, (iii) dynamics of membrane potentials, and (iv) arbitrary fluxes and passive cross-membrane transport mechanisms, and supports PDEs in addition to ODEs. It is one of very few tools to incorporate physicochemical and electro-physiological processes and can apply quasi-steady-state approximations to fast reactions. It is also an image processing tool for experimental images.

The simulation environment MOOSE was developed as a reimplementation of the GENESIS neural simulator, and initially used that simulator's model description format. Recently though it has developed support for NeuroML models, and is also capable of dealing with systems biological models (Gleeson et al., 2010; Dudani et al., 2013). New simulation algorithms can be added to MOOSE through a generic framework. It has also been developed with a focus on multi-scale models and simulation in diverse levels of detail (Dudani et al., 2013). For a more comprehensive overview about recent simulation tools with a focus on neuroscience we refer the interested reader to the review by Gleeson (2013).

The stand-alone modeling framework PhysioDesigner (Asai et al., 2013) provides several functions for the creation and analysis of PHML models. SBML models can be incorporated as submodels through PhysioDesigner (Asai et al., 2014), aiming at integrating dynamics at sub-cellular and cellular levels. The simulator Flint can efficiently solve PHML models and provides a cloud service, which allows users to remotely solve their models (Asai et al., 2012). PhysioDesigner uses Flint and submits jobs to this cloud service.

3.2.4. *Regulatory networks*

The inference of regulatory networks is a challenge for many areas of research. The program ModuleMaster (Wrzodek et al., 2010) identifies *cis*-regulatory modules (CRMs) in sets of co-expressed genes based on transcription factor binding information and multivariate functional relationships between regulators and target genes. As an input it uses microarray and clustering experiments and SBML models as output. In order to make the results of network inference procedures such as Net*Gene*rator (Töpfer et al., 2007) reusable in further analysis tools, the program GRN2SBML (Vlaic et al., 2013) has been developed as a converter to SBML. It provides a graphical user interface, access to BioMart Central, and can also be used as an R-package. The program GINsim has been developed for the analysis and simulation of logical models of gene interaction networks (Gonzalez Gonzalez et al., 2006) and has been recently adapted to the SBML qual extension (Chaouiya et al., 2013). The program CellNOpt can be useful for the creation of signal transduction networks based on a logical approach (Terfve et al., 2012), and it also supports SBML qual.

3.3. REGULAR COMMUNITY MEETINGS

Many standards described in this paper are based on community efforts. For this reason, community meetings have been required from their inception. In October 2010, separate workshops were combined in order to better coordinate individual developments and to reduce the necessary amount of traveling for individual researchers. This resulted in two regular annual meetings that brought together the community. The COMBINE (Computational Modeling in Biology Network) is a workshop with scientific presentations, poster sessions, and several break-out sessions, which are used to discuss and coordinate the further development of the "COMBINE Standards" BioPAX, CellML, SBGN, SBML, SBOL, and SED-ML, as well as associated and related standards. The idea of the spring Hackathon on resources for modeling in biology (HARMONY) is to provide room and time for community members to sit down, share code and ideas, program, and discuss. In contrast to the fall event, HARMONY usually has only very few talks and is much more a hands-on practical event, where participants develop new approaches and ideas. For more information about previous meetings see the meeting reports by Le Novère et al. (2011), Waltemath et al. (2014) and the COMBINE homepage[16]. This alternating sequence of complementary meetings leads to a very efficient and progressive development of software and standards.

4. DISCUSSION

In this review article, we have examined diverse modeling standards and data formats that are currently in use within the scientific community together, with databases from where these formats can be obtained. We discussed a selection of useful software packages and modeling approaches for systems biology and related fields. The structuring of individual standards is at present very elaborate: there is usually a modeling, annotation, or documentation recommendation that forms the theoretical basis for a corresponding machine-readable data format and involves specific controlled vocabulary terms for unambiguous specification of individual model components.

Aiming to keep even highly elaborate standards flexible and able to incorporate new findings, the specifications are becoming more and more abstract and modularized. For example, the original *reaction* element in SBML is now seen as a generic *process* whose inputs and outputs no longer strictly have to *represent* substrates and products of biochemical reactions. The idea to develop specific packages for certain needs rather than one monolithic modeling language also follows this trend. The development of all standards involves numerous people, detailed discussions, and careful consideration. This overall procedure ensures that standards mature in an open fashion and allows interested researchers to participate and to contribute to this development. At the same time, it also increases the chance that potential conflicts or inaccuracies can be discovered in early stages of development. With increased use of standards the requirements of the individual format are steadily improved and current limitations are detected and solved. Thanks to the regular meetings and ongoing exchange between the developers of the diverse standards, the individual formats are mutually

[16] http://co.mbine.org

adopting more and more of each other's features. It can therefore be expected that the exchange between different model and pathway representation standards will further increase.

For end-user applications, the goal is that users would no longer have to care about the underlying data format used by a specific software tool. More and more details of the internal structure and organization of underlying formats could be hidden and no detailed knowledge about these formats would be required. Plug-ins for platforms such as Cytoscape (Shannon et al., 2003) or CellDesigner (Funahashi et al., 2008) can provide complementary functionality for export or import of certain data formats based on a common underlying data structure (König et al., 2012; Gonçalves et al., 2013). The SBW or the Garuda framework provides further ways to increase the interoperability of tools with little effort (Sauro et al., 2003; Ghosh et al., 2011). Many tools could also benefit from the ability of the new COMBINE archive format to bridge separately stored representations or applications of the same model (Bergmann et al., 2014).

The distribution and curation of standardized models, their simulation description, and expected results by centralized databases plays a prominent role. These knowledge bases constitute valuable resources of available information about biological processes and reproducible experiments. They can therefore significantly reduce time and effort needed for the assembly of extended models and create the basis for further research. The ability to easily reproduce new scientific findings with existing simulation workflows facilitates the fast adoption and integration of these findings into new and even further elaborated works. If other researchers are able to run simulations and to comprehend models with minimal effort, it can be expected that these studies will receive higher recognition and lead to more citations compared to distributing models whose outcomes are difficult to reproduce. The distribution of models and data in standard formats amongst their project working groups will not only benefit collaboration partners, but the fine-grained structure of standards for diverse aspects of modeling workflows that is now available can even simplify the review process of scientific papers. If a model is uploaded along with a publication in a standard format, accompanied with a simulation experiment description file and a graphical representation, reviewers can quickly obtain an overview about structure and organization of a model, and even easily check if the findings described in the paper can be reproduced. Thereby, the reviewer can select any numerical tool that supports these data formats and is not restricted to any particular environment.

The development of a standard can be seen as a long-term investment. Unlike in other fields, the community-based bottom-up development of exchange formats is very common in systems biology. Depending on the structure of the field, it can therefore take a long time before the overhead of developing a new standard pays off; on the other hand, standards exist as long as the community has a requirement for them (Brazma et al., 2006). It also seems that the development of standards has become a field of research by itself and is sometimes even seen as the central aspect in modeling (Waltemath et al., 2013). Models and their evaluation are certainly valuable tools for progress in research, but permanently keeping track of all emerging standards can become difficult. The proposed concepts and approaches can only be successful if these are well-known. If standard data formats are developed that are not adopted by the community, the standard will disappear and a simpler solution will gain acceptance. As we go along, new modeling techniques and new finding are established and adopted by the research community (Lerman et al., 2012; O'Brien et al., 2013). Approaches for model encoding and standardization therefore need to continuously evolve with the domain of research that they represent. It is therefore important for the standardization community to continue to closely interact with the modeling community in order to catch up with novel approaches, needs, and requirements. The solutions given to the modeling community must be simple enough in order to be easily adopted, implemented, and applied, but they must also be sophisticated enough in order to capture the complexity of the described systems. Participation of the community in proposing encoding schemes and guideline checklists is essential for the success of the respective standard. Large-scale reconstructions and community projects require data standards and at the same time push their development (Büchel et al., 2013a; Thiele et al., 2013).

While in the past even quick computation in active research required the implementation of some data structures from scratch in customized scripts, the rich variety of software libraries and modeling-specific scripting languages now available drastically simplify these tasks. If an existing software solution cannot be directly applied to solve a specific task, it is at least possible to use standards compliant data structures from the very beginning of a project. Also the quality of available software solutions is progressively increasing. For the distribution of final results, standard formats should be used as the preferred exchange and storage medium in order to ensure reusability and reproducibility of results and findings.

AUTHOR CONTRIBUTIONS

Andreas Dräger developed the conceptual idea for this review and drafted the manuscript. Bernhard Ø. Palsson supervised this work and critically revised the manuscript. Both authors approve the final manuscript.

ACKNOWLEDGMENTS

The authors thank Marc Abrahams, UCSD, for carefully proofreading the manuscript and acknowledge the contribution of Nikolaus Sonnenschein, Novo Nordisk Foundation Center for Biosustainability (CFB). *Funding*: This work was funded by a Marie Curie International Outgoing Fellowship within the European Commission's 7th Framework Program for Research and Technological Development (project AMBiCon, grant number 332020) and a National Institute of Health grant for the continued development of essential SBML software support (NIH, United States, award number GM070923).

REFERENCES

Adams, R. R. (2012). SED-ED, a workflow editor for computational biology experiments written in SED-ML. *Bioinformatics* 28, 1180–1181. doi:10.1093/bioinformatics/bts101

Asai, Y., Abe, T., Oka, H., Okita, M., Hagihara, K.-I., Ghosh, S., et al. (2014). A versatile platform for multilevel modeling of physiological systems: SBML-PHML hybrid modeling and simulation. *Adv Biomed Eng* 3, 50–58. doi:10.1109/EMBC.2013.6610802

Asai, Y., Abe, T., Oka, H., Okita, M., Okuyama, T., Hagihara, K.-I., et al. (2013). "A versatile platform for multilevel modeling of physiological systems: template/instance framework for large-scale modeling and simulation," in *35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. Osaka: IEEE.

Asai, Y., Abe, T., Okita, M., Okuyama, T., Yoshioka, N., Yokoyama, S., et al. (2012). "Multilevel modeling of physiological systems and simulation platform: physiodesigner, flint and flint K3 service," in *IEEE/IPSJ 12th International Symposium on Applications and the Internet*. Izmir: IEEE.

Asai, Y., Suzuki, Y., Kido, Y., Oka, H., Heien, E., Nakanishi, M., et al. (2008). Specifications of insilicoML 1.0: a multilevel biophysical model description language. *J. Physiol. Sci.* 58, 447–458. doi:10.2170/physiolsci.RP013308

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., et al. (2000). Gene ontology: tool for the unification of biology. The gene ontology consortium. *Nat. Genet.* 25, 25–29. doi:10.1038/75556

Beazley, D. M. (1996). *SWIG: An Easy to Use Tool for Integrating Scripting Languages with C and C++. Technical Report.* Salt Lake City, UT: Department of Computer Science, University of Utah.

Becker, S. A., Feist, A. M., Mo, M. L., Hannum, G., Palsson, B. Ø., and Herrgård, M. J. (2007). Quantitative prediction of cellular metabolism with constraint-based models: the COBRA toolbox. *Nat. Protoc.* 2, 727–738. doi:10.1038/nprot.2007.99

Beeman, D. (2013). "History of neural simulation software," in *Years of Computational Neuroscience*, , Vol. 9, ed. J. M. Bower (New York: Springer), 33–71.

Bergmann, F. T., Adams, R., Moodie, S., Cooper, J., Glont, M., Golebiewski, M., et al. (2014). *Combine Archive: One File to Share Them All.* Ithaca: Cornell University Library.

Bergmann, F. T., and Olivier, B. G. (2013). "SBML level 3 package: flux balance constraints ('fbc')," in *Technical Report.* Pasadena, CA: Caltech. Available from: http://co.mbine.org/specifications/sbml.level-3.version-1.fbc.version-1.release-1.pdf

Bergmann, F. T., and Sauro, H. M. (2008). Comparing simulation results of SBML capable simulators. *Bioinformatics* 24, 1963–1965. doi:10.1093/bioinformatics/btn319

Booth, I. R. (2007). Sysmo: back to the future. *Nat. Rev. Microbiol.* 5, 566–566. doi:10.1038/nrmicro1719

Bordbar, A., Monk, J. M., King, Z. A., and Palsson, B. Ø. (2014). Constraint-based models predict metabolic and associated cellular functions. *Nat. Rev. Gen.* 15, 107–120. doi:10.1038/nrg3643

Bornstein, B. J., Keating, S. M., Jouraku, A., and Hucka, M. (2008). LibSBML: an API library for SBML. *Bioinformatics* 24, 880–881. doi:10.1093/bioinformatics/btn051

Brazma, A., Krestyaninova, M., and Sarkans, U. (2006). Standards for systems biology. *Nat Rev Genet* 7, 593–605. doi:10.1038/nrg1922

Britten, R. D., Christie, G. R., Little, C., Miller, A. K., Bradley, C., Wu, A., et al. (2013). FieldML, a proposed open standard for the physiome project for mathematical model representation. *Med. Biol. Eng. Comput.* 51, 1191–1207. doi:10.1007/s11517-013-1097-7

Brooksbank, C., Bergman, M. T., Apweiler, R., Birney, E., and Thornton, J. (2013). The European bioinformatics institute's data resources 2014. *Nucleic Acids Res.* 42, D18–D25. doi:10.1093/nar/gkt1206

Büchel, F., Rodriguez, N., Swainston, N., Wrzodek, C., Czauderna, T., Keller, R., et al. (2013a). Large-scale generation of computational models from biochemical pathway maps. *BMC Syst. Biol.* 7:116. doi:10.1186/1752-0509-7-116

Büchel, F., Saliger, S., Dräger, A., Hoffman, S., Wrzodek, C., Zell, A., et al. (2013b). Parkinson's disease: dopaminergic nerve cell model is consistent with experimental finding of increased extracellular transport of α-synuclein. *BMC Neurosci.* 14:136. doi:10.1186/1471-2202-14-136

Büchel, F., Wrzodek, C., Mittag, F., Dräger, A., Eichner, J., Rodriguez, N., et al. (2012). Qualitative translation of relations from BioPAX to SBML qual. *Bioinformatics* 28, 2648–2653. doi:10.1093/bioinformatics/bts508

Butterworth, E., Jardine, B. E., Raymond, G. M., Neal, M. L., and Bassingthwaighte, J. B. (2014). JSim, an open-source modeling system for data analysis. *F1000Res.* 2:288. doi:10.12688/f1000research.2-288.v3

Cannon, R. C., Gleeson, P., Crook, S., Ganapathy, G., Marin, B., Piasini, E., et al. (2014). LEMS: a language for expressing complex biological models in concise and hierarchical form and its use in underpinning NeuroML 2. *Front. Neuroinform.* 8:79. doi:10.3389/fninf.2014.00079

Caspi, R., Altman, T., Billington, R., Dreher, K., Foerster, H., Fulcher, C. A., et al. (2014). The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. *Nucleic Acids Res.* 42, D459–D471. doi:10.1093/nar/gkt1103

Chandran, D., Bergmann, F. T., and Sauro, H. M. (2009). TinkerCell: modular CAD tool for synthetic biology. *J. Biol. Eng.* 3, 19. doi:10.1186/1754-1611-3-19

Chaouiya, C., Berenguier, 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

Chelliah, V., Laibe, C., and Le Novère, N. (2013). "BioModels database: a repository of mathematical models of biological processes," in *In silico Systems Biology*, volume 1021 of *Methods in Molecular Biology*, ed. M. V. Schneider (New York: Springer), 189–199.

Cooling, M. T. (2010). "A primer on modular mass-action modelling with CellML," in *Systems Biology for Signaling Networks*, volume 1 of *Systems Biology*, ed. S. Choi (New York: Springer), 721–750.

Cooling, M. T., Rouilly, V., Misirli, G., Lawson, J., Yu, T., Hallinan, J., et al. (2010). Standard virtual biological parts: a repository of modular modeling components for synthetic biology. *Bioinformatics* 26, 925–931. doi:10.1093/bioinformatics/btq063

Courtot, M., Juty, N., Knüpfer, C., Waltemath, D., Zhukova, A., Dräger, A., et al. (2011). Controlled vocabularies and semantics in systems biology. *Mol. Syst. Biol.* 7, 543. doi:10.1038/msb.2011.77

Crasto, C. J., Marenco, L. N., Liu, N., Morse, T. M., Cheung, K.-H., Lai, P. C., et al. (2007). SenseLab: new developments in disseminating neuroscience information. *Brief. Bioinformatics* 8, 150–162. doi:10.1093/bib/bbm018

Cuellar, A., Nielsen, P., Halstead, M., Bullivant, D., Nickerson, D., Hedley, W., et al. (2006). *CellML 1.1 Specification. Technical report.* Auckland, NZ: Bioengineering Institute, University of Auckland.

Czauderna, T., Klukas, C., and Schreiber, F. (2010). Editing, validating and translating of SBGN maps. *Bioinformatics* 26, 2340–2341. doi:10.1093/bioinformatics/btq407

Czauderna, T., Wybrow, M., Marriott, K., and Schreiber, F. (2013). Conversion of KEGG metabolic pathways to SBGN maps including automatic layout. *BMC Bioinformatics* 14:250. doi:10.1186/1471-2105-14-250

Dada, J. O., Spasic, I., Paton, N. W., and Mendes, P. (2010). SBRML: a markup language for associating systems biology data with models. *Bioinformatics* 26, 932–938. doi:10.1093/bioinformatics/btq069

Dandekar, T., Fieselmann, A., Majeed, S., and Ahmed, Z. (2012). Software applications toward quantitative metabolic flux analysis and modeling. *Brief. Bioinformatics* 15, 91–107. doi:10.1093/bib/bbs065

Deckard, A., Bergmann, F. T., and Sauro, H. M. (2006). Supporting the SBML layout extension. *Bioinformatics* 22, 2966–2967. doi:10.1093/bioinformatics/btl520

Demir, E., Babur, Ö, Rodchenkov, I., Aksoy, B. A., Fukuda, K. I., Gross, B., et al. (2013). Using biological pathway data with paxtools. *PLoS Comput. Biol.* 9:e1003194. doi:10.1371/journal.pcbi.1003194

Demir, E., Cary, M. P., Paley, S., Fukuda, K., Lemer, C., Vastrik, I., et al. (2010). The BioPAX community standard for pathway data sharing. *Nat. Biotechnol.* 28, 935–942. doi:10.1038/nbt.1666

Dräger, A. (2011). *Computational Modeling of Biochemical Networks.* München: Eberhard Karls University of Tübingen.

Dräger, A., Hassis, N., Supper, J., Schröder, A., and Zell, A. (2008). SBMLsqueezer: a CellDesigner plug-in to generate kinetic rate equations for biochemical networks. *BMC Syst. Biol.* 2:39. doi:10.1186/1752-0509-2-39

Dräger, A., Kronfeld, M., Ziller, M. J., Supper, J., Planatscher, H., Magnus, J. B., et al. (2009a). Modeling metabolic networks in *C. glutamicum*: a comparison of rate laws in combination with various parameter optimization strategies. *BMC Syst. Biol.* 3:5. doi:10.1186/1752-0509-3-5

Dräger, A., Planatscher, H., Wouamba, D. M., Schröder, A., Hucka, M., Endler, L., et al. (2009b). SBML2L$_{A}$T$_{E}$X: conversion of SBML files into human-readable reports. *Bioinformatics* 25, 1455–1456. doi:10.1093/bioinformatics/btp170

Dräger, A., and Planatscher, H. (2013a). *Encyclopedia of Systems Biology*, Chapter Metabolic Networks. New York, Heidelberg, Dordrecht, London: Springer-Verlag, 1249–1251.

Dräger, A., and Planatscher, H. (2013b). *Encyclopedia of Systems Biology*, Chapter Parameter Estimation, Metabolic Network Modeling. New York, Heidelberg, Dordrecht, London: Springer-Verlag, 1627–1631.

Dräger, A., Rodriguez, N., Dumousseau, M., Drr, A., Wrzodek, C., Le Novère, N., et al. (2011). JSBML: a flexible Java library for working with SBML. *Bioinformatics* 27, 2167–2168. doi:10.1093/bioinformatics/btr361

Dräger, A., Schröder, A., and Zell, A. (2010). *Systems Biology for Signaling Networks*, volume 2, chapter Automating Mathematical Modeling of Biochemical Reaction Networks. New York: Springer.

Duarte, N. C., Becker, S. A., Jamshidi, N., Thiele, I., Mo, M. L., Vo, T. D., et al. (2007). Global reconstruction of the human metabolic network based on genomic and bibliomic data. *Proc. Natl. Acad. Sci. U. S. A.* 104, 1777–1782. doi:10.1073/pnas.0610772104

Dudani, N., Bhalla, U. S., and Ray, S. (2013). MOOSE, the multiscale object-oriented simulation environment. *Encyclopedia of Computational Neuroscience* 1–4. doi:10.3389/neuro.11.006.2008

Ebrahim, A., Lerman, J. A., Palsson, B. Ø., and Hyduke, D. R. (2013). COBRApy: constraints-based reconstruction and analysis for python. *BMC Syst. Biol.* 7:74. doi:10.1186/1752-0509-7-74

Endler, L., Rodriguez, N., Juty, N., Chelliah, V., Laibe, C., Li, C., et al. (2009). Designing and encoding models for synthetic biology. *J. R. Soc. Interface* 6(Suppl. 4), S405–S417. doi:10.1098/rsif.2009.0035.focus

Feist, A. M., Henry, C. S., Reed, J. L., Krummenacker, M., Joyce, A. R., Karp, P. D., et al. (2007). A genome-scale metabolic reconstruction for escherichia coli k-12 mg1655 that accounts for 1260 orfs and thermodynamic information. *Mol. Syst. Biol.* 3, 121. doi:10.1038/msb4100155

Fernández-Castané, A., Fehér, T., Carbonell, P., Pauthenier, C., and Faulon, J.-L. (2014). Computer-aided design for metabolic engineering. *J. Biotechnol.* doi:10.1016/j.jbiotec.2014.03.029

Finney, A., and Hucka, M. (2003). Systems biology markup language: level 2 and beyond. *Biochem. Soc. Trans.* 31(Pt 6), 1472–1473. doi:10.1042/BST0311472

Funahashi, A., Matsuoka, Y., Jouraku, A., Morohashi, M., Kikuchi, N., and Kitano, H. (2008). "CellDesigner 3.5: a versatile modeling tool for biochemical networks," in *Proceedings of the IEEE*, Vol. 96 (IEEE), 1254–1265.

Galdzicki, M., Clancy, K. P., Oberortner, E., Pocock, M., Quinn, J. Y., Rodriguez, C. A., et al. (2014). The synthetic biology open language (SBOL) provides a community standard for communicating designs in synthetic biology. *Nat. Biotechnol.* 32, 545–550. doi:10.1038/nbt.2891

Galdzicki, M., Rodriguez, C., Chandran, D., Sauro, H. M., and Gennari, J. H. (2011). Standard biological parts knowledgebase. *PLoS ONE* 6:e17005. doi:10.1371/journal.pone.0017005

Gauges, R., Rost, U., Sahle, S., and Wegner, K. (2006). A model diagram layout extension for SBML. *Bioinformatics* 22, 1879–1885. doi:10.1093/bioinformatics/btl195

Gerasch, A., Faber, D., Küntzer, J., Niermann, P., Kohlbacher, O., Lenhof, H.-P., et al. (2014). Bina: a visual analytics tool for biological network data. *PLoS ONE* 9:e87397. doi:10.1371/journal.pone.0087397

Ghosh, S., Matsuoka, Y., Asai, Y., Hsin, K.-Y., and Kitano, H. (2011). Software for systems biology: from tools to integrated platforms. *Nat Rev Genet* 12, 821–832. doi:10.1038/nrg3096

Gleeson, P. (2013). "Software tools for modelling in computational neuroscience: overview," in *Encyclopedia of Computational Neuroscience*, eds D. Jaeger and R. Jung (New York: Springer), 1–4. doi:10.1007/978-1-4614-7320-6_93-1

Gleeson, P., Crook, S., Cannon, R. C., Hines, M. L., Billings, G. O., Farinella, M., et al. (2010). NeuroML: a language for describing data driven models of neurons and networks with a high degree of biological detail. *PLoS Comput. Biol.* 6:e1000815. doi:10.1371/journal.pcbi.1000815

Gleeson, P., Silver, R. A., and Cantarelli, M. (2013). "Open source brain", in *Encyclopedia of Computational Neuroscience*, eds D. Jaeger and R. Jung (New York: Springer), 1–3. doi:10.1007/978-1-4614-7320-6_595-2

Goddard, N. H., Hucka, M., Howell, F., Cornelis, H., Shankar, K., and Beeman, D. (2001). Towards NeuroML: model description methods for collaborative modelling in neuroscience. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 356, 1209–1228. doi:10.1098/rstb.2001.0910

Gonçalves, E., van Iersel, M., and Saez-Rodriguez, J. (2013). CySBGN: a cytoscape plug-in to integrate SBGN maps. *BMC Bioinformatics* 14:17. doi:10.1186/1471-2105-14-17

Gonzalez Gonzalez, A., Naldi, A., Sánchez, L., Thieffry, D., and Chaouiya, C. (2006). GINsim: a software suite for the qualitative modelling, simulation and analysis of regulatory networks. *BioSystems* 84, 91–100. doi:10.1016/j.biosystems.2005.10.003

Gostner, R., Baldacci, B., Morine, M. J., and Priami, C. (2014). Graphical modeling tools for systems biology. *ACM Comput. Surv.* 47, 1–21. doi:10.1145/2633461

Grafahrend-Belau, E., Klukas, C., Junker, B. H., and Schreiber, F. (2009). FBA-SimVis: interactive visualization of constraint-based metabolic models. *Bioinformatics* 25, 2755–2757. doi:10.1093/bioinformatics/btp408

Grillner, S. (2014). Megascience efforts and the brain. *Neuron* 82, 1209–1211. doi:10.1016/j.neuron.2014.05.045

Hamilton, J. J., and Reed, J. L. (2013). Software platforms to facilitate reconstructing genome-scale metabolic networks. *Environ. Microbiol.* 16, 49–59. doi:10.1111/1462-2920.12312

Hastings, J., de Matos, P., Dekker, A., Ennis, M., Harsha, B., Kale, N., et al. (2013). The ChEBI reference database and ontology for biologically relevant chemistry: enhancements for 2013. *Nucleic Acids Res.* 41, D456–D463. doi:10.1093/nar/gks1146

Henkel, R., Endler, L., Peters, A., Le Novre, N., and Waltemath, D. (2010). Ranked retrieval of computational biology models. *BMC Bioinformatics* 11:423. doi:10.1186/1471-2105-11-423

Herrgård, M. J., Swainston, N., Dobson, P., Dunn, W. B., Arga, K. Y., Arvas, M., et al. (2008). A consensus yeast metabolic network reconstruction obtained from a community approach to systems biology. *Nat. Biotechnol.* 26, 1155–1160. doi:10.1038/nbt1492

Hindmarsh, A. C., Brown, P. N., Grant, K. E., Lee, S. L., Serban, R., Shumaker, D. E., et al. (2005). SUNDIALS: suite of nonlinear and differential/algebraic equation solvers. *ACM T Math Software* 31, 363–396. doi:10.1145/1089014.1089020

Hines, M. L., Morse, T., Migliore, M., Carnevale, N. T., and Hines, M. L. (2004). ModelDB: a database to support computational neuroscience. *J. Comput. Neurosci.* 17, 7–11. doi:10.1023/B:JCNS.0000023869.22017.2e

Holzhütter, H.-G., Drasdo, D., Preusser, T., Lippert, J., and Henney, A. M. (2012). The virtual liver: a multidisciplinary, multilevel challenge for systems biology. *Wiley Interdiscip Rev Syst Biol Med* 4, 221–235. doi:10.1002/wsbm.1158

Hoops, S., Sahle, S., Gauges, R., Lee, C., Pahle, J., Simus, N., et al. (2006). Copasi – a complex pathway simulator. *Bioinformatics* 22, 3067–3074. doi:10.1093/bioinformatics/btl485

Hoppe, A., Hoffmann, S., Gerasch, A., Gille, C., and Holzhütter, H.-G. (2011). FASIMU: flexible software for flux-balance computation series in large metabolic networks. *BMC Bioinformatics* 12:28. doi:10.1186/1471-2105-12-28

Hucka, M., Finney, A., Bornstein, B. J., Keating, S. M., Shapiro, B. E., Matthews, J., et al. (2004). Evolving a lingua franca and associated software infrastructure for computational systems biology: the systems biology markup language (SBML) project. *Syst Biol (Stevenage)* 1, 41–53. doi:10.1049/sb:20045008

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

Hunter, P. J., and Borg, T. K. (2003). Integration from proteins to organs: the physiome project. *Nat. Rev. Mol. Cell Biol.* 4, 237–243. doi:10.1038/nrm1054

Junker, B. H., Klukas, C., and Schreiber, F. (2006). VANTED: a system for advanced data analysis and visualization in the context of biological networks. *BMC Bioinformatics* 7:219. doi:10.1186/1471-2105-7-219

Juty, N., Laibe, C., and Le Novère, N. (2013). "Controlled annotations for systems biology," in *In silico Systems Biology*, Vol. 1021, ed. M. V. Schneider (New York: Springer), 227–245.

Juty, N., Le Novère, N., and Laibe, C. (2012). Identifiers.org and MIRIAM registry: community resources to provide persistent identification. *Nucleic Acids Res.* 40, D580–D586. doi:10.1093/nar/gkr1097

Kandel, E. R., Markram, H., Matthews, P. M., Yuste, R., and Koch, C. (2013). Neuroscience thinks big (and collaboratively). *Nat. Rev. Neurosci.* 14, 659–664. doi:10.1038/nrn3578

Kanehisa, M., and Goto, S. (2000). KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* 28, 27–30. doi:10.1093/nar/28.1.27

Keating, S. M., Bornstein, B. J., Finney, A., and Hucka, M. (2006). SBMLToolbox: an SBML toolbox for MATLAB users. *Bioinformatics* 22, 1275–1277. doi:10.1093/bioinformatics/btl111

Kelder, T., van Iersel, M. P., Hanspers, K., Kutmon, M., Conklin, B. R., Evelo, C. T., et al. (2011). WikiPathways: building research communities on biological pathways. *Nucleic Acids Res.* 40, D1301–D1307. doi:10.1093/nar/gkr1074

Keller, R., Dörr, A., Tabira, A., Funahashi, A., Ziller, M. J., Adams, R., et al. (2013). The systems biology simulation core algorithm. *BMC Syst. Biol.* 7:55. doi:10.1186/1752-0509-7-55

Kitano, H., Funahashi, A., Matsuoka, Y., and Oda, K. (2005). Using process diagrams for the graphical representation of biological networks. *Nat. Biotechnol.* 23, 961–966. doi:10.1038/nbt1111

Knüpfer, C., Beckstein, C., and Dittrich, P. (2006). "Towards a semantic description of biomodels: meaning facets – a case study," in *Proceedings of the Second International Symposium on Semantic Mining in Biomedicine (SMBM 2006)*, eds S. Ananiadou, S. Pyysalo, D. Rebholz-Schuhmann, F. Rinaldi, and T. Salakoski (Jena: CEUR-WS), 97–100.

Kohn, K. W., Aladjem, M. I., Weinstein, J. N., and Pommier, Y. (2006). Molecular interaction maps of bioregulatory networks: a general rubric for systems biology. *Mol. Biol. Cell* 17, 1–13. doi:10.1091/mbc.E05-09-0824

Kolpakov, F. A., Tolstykh, N. I., Valeev, T. F., Kiselev, I. N., Kutumova, E. O., Ryabova, A., et al. (2011). "BioUML-open source plug-in based platform for bioinformatics: invitation to collaboration," in *Moscow Conference on Computational Molecular Biology* (Moskow: Department of Bioengineering and Bioinformatics of MV Lomonosov Moscow State University), 172–173.

König, M., Dräger, A., and Holzhütter, H.-G. (2012). CySBML: a cytoscape plugin for SBML. *Bioinformatics* 28, 2402–2403. doi:10.1093/bioinformatics/bts432

Koussa, J., Chaiboonchoe, A., and Salehi-Ashtiani, K. (2014). Computational approaches for microalgal biofuel optimization: a review. *Biomed Res. Int.* 2014, 1–12. doi:10.1155/2014/649453

Kramer, F., Bayerlová, M., and Beißbarth, T. (2014). R-based software for the integration of pathway data into bioinformatic algorithms. *Biology* 3, 85–100. doi:10.3390/biology3010085

Krause, F., Schulz, M., Ripkens, B., Flöttmann, M., Krantz, M., Klipp, E., et al. (2013). Biographer: web-based editing and rendering of SBGN compliant biochemical networks. *Bioinformatics* 29, 1467–1468. doi:10.1093/bioinformatics/btt159

Kronfeld, M. (2008). *EvA2 Short Documentation.* Tübingen: University of Tübingen, Department of Computer Architecture.

Kronfeld, M., Dräger, A., Aschoff, M., and Zell, A. (2009). "On the benefits of multimodal optimization for metabolic network modeling," in *German Conference on Bioinformatics (GCB 2009)*, Volume P-157 of *Lecture Notes in Informatics*, eds I. Grosse, S. Neumann, S. Posch, F. Schreiber, and P. Stadler (Halle: German Informatics society), 191–200.

Küntzer, J., Backes, C., Blum, T., Gerasch, A., Kaufmann, M., Kohlbacher, O., et al. (2007). Bndb – the biochemical network database. *BMC Bioinformatics* 8:367. doi:10.1186/1471-2105-8-367

Laible, C., and Le Novère, N. (2007). MIRIAM resources: tools to generate and resolve robust cross-references in systems biology. *BMC Syst. Biol.* 13:58. doi:10.1186/1752-0509-1-58

Lambeck, S., Dräger, A., and Guthke, R. (2010). "Network inference by considering multiple objectives: insights from in vivo transcriptomic data generated by a synthetic network," in *International Conference on Bioinformatics and Computational Biology, BIOCOMP 2010*, Vol. 2, eds H. R. Arabnia, Q.-N. Tran, R. Chang, M. He, A. Marsh, A. M. G. Solo, et al. (Las Vegas, NV: CSREA Press), 734–742.

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. doi:10.1038/nbt1156

Le Novère, N., Hucka, M., Anwar, N., Bader, G. D., Demir, E., Moodie, S., et al. (2011). Meeting report from the first meetings of the computational modeling in biology network (combine). *Stand. Genomic Sci.* 5, 230–242. doi:10.4056/sigs.2034671

Le Novère, N., Hucka, M., Mi, H., Moodie, S., Schreiber, F., Sorokin, A., et al. (2009). The systems biology graphical notation. *Nat. Biotechnol.* 27, 735–741. doi:10.1038/nbt.1558

Le Novere, N., Le Novere, N., Demir, E., Mi, H., Moodie, S., and Villeger, A. (2011). Systems biology graphical notation: entity relationship language level 1 (version 1.2). *Nature Precedings* doi:10.1038/npre.2011.5902.1

Lerman, J. A., Hyduke, D. R., Latif, H., Portnoy, V. A., Lewis, N. E., Orth, J. D., et al. (2012). In silico method for modelling metabolism and gene product expression at genome scale. *Nat. Commun.* 3, 929. doi:10.1038/ncomms1928

Lloyd, C. (2013). Opencell. *Encyclopedia of Systems Biology* 1567–1568. doi:10.1007/978-1-4419-9863-7_1526

Lloyd, C. M., Halstead, M. D. B., and Nielsen, P. F. (2004). CellML: its future, present and past. *Prog Biophys Mol Biol* 85, 433–450. doi:10.1016/j.pbiomolbio.2004.01.004

Machné, R., Finney, A., Müller, S., Lu, J., Widder, S., and Flamm, C. (2006). The SBML ODE solver library: a native API for symbolic and fast numerical analysis of reaction networks. *Bioinformatics* 22, 1406–1407. doi:10.1093/bioinformatics/btl086

Macilwain, C. (2011). Systems biology: evolving into the mainstream. *Cell* 144, 839–841. doi:10.1016/j.cell.2011.02.044

Markram, H., Meier, K., Lippert, T., Grillner, S., Frackowiak, R., Dehaene, S., et al. (2011). Introducing the human brain project. *Procedia Compu Sci* 7, 39–42. doi:10.1016/j.procs.2011.12.015

Matsuoka, Y., Funahashi, A., Ghosh, S., and Kitano, H. (2014). Modeling and simulation using CellDesigner. *Methods Mol. Biol.* 1164, 121–145. doi:10.1007/978-1-4939-0805-9_11

Mi, H., Muruganujan, A., Demir, E., Matsuoka, Y., Funahashi, A., Kitano, H., et al. (2011). BioPAX support in CellDesigner. *Bioinformatics* 27, 3437–3438. doi:10.1093/bioinformatics/btr586

Migliore, M., Morse, T. M., Davison, A. P., Marenco, L., Shepherd, G. M., and Hines, M. L. (2003). ModelDB: making models publicly accessible to support computational neuroscience. *Neuroinformatics* 1, 135–140. doi:10.1385/NI:1:1:135

Miller, A. K., Britten, R. D., and Nielsen, P. M. F. (2012). Declarative representation of uncertainty in mathematical models. *PLoS ONE* 7:e39721. doi:10.1371/journal.pone.0039721

Miller, A. K., Marsh, J., Reeve, A., Garny, A., Britten, R., Halstead, M., et al. (2010). An overview of the cellml api and its implementation. *BMC Bioinformatics* 11:178. doi:10.1186/1471-2105-11-178

Moodie, S. L., Swat, M. J., Kristensen, N. R., and Le Novère, N. (2013). PharmML: the pharmacometrics markup language. Available from: https://sites.google.com/site/pharmmltemp/documentation2/pharmml-specification_0.2.1.pdf

Moraru, I. I., Schaff, J. C., Slepchenko, B. M., Blinov, M. L., Morgan, F., Lakshminarayana, A., et al. (2008). Virtual cell modelling and simulation software environment. *IET Syst. Biol.* 2, 352–362. doi:10.1049/iet-syb:20080102

Müller, K. M., and Arndt, K. M. (2012). "Standardization in synthetic biology," in *Synthetic Gene Networks*, eds W. Weber and M. Fussenegger (New York: Springer), 23–43.

Myers, C. J., Barker, N., Jones, K., Kuwahara, H., Madsen, C., and Nguyen, N.-P. D. (2009). ibiosim: A tool for the analysis and design of genetic circuits. *Bioinformatics* 25, 2848–2849. doi:10.1093/bioinformatics/btp457

Nickerson, D. P., Garny, A., Nielsen, P. M. F., and Hunter, P. J. (2013). "Standards and tools supporting collaborative development of the virtual physiological human," in *Engineering in Medicine and Biology Society (EMBC), 2013 35th Annual International Conference of the IEEE* (Osaka: IEEE), 5541–5544.

O'Brien, E. J., Lerman, J. A., Chang, R. L., Hyduke, D. R., and Palsson, B. Ø. (2013). Genome-scale models of metabolism and gene expression extend and refine growth phenotype prediction. *Mol. Syst. Biol.* 9, 693. doi:10.1038/msb.2013.52

Olivier, B. G., Rohwer, J. M., and Hofmeyr, J.-H. S. (2005). Modelling cellular systems with PySCeS. *Bioinformatics* 21, 560–561. doi:10.1093/bioinformatics/bti046

Olivier, B. G., and Snoep, J. L. (2004). Web-based kinetic modelling using JWS Online. *Bioinformatics* 20, 2143–2144. doi:10.1093/bioinformatics/bth200

Orth, J. D., Thiele, I., and Palsson, B. Ø. (2010). What is flux balance analysis? *Nat. Biotechnol.* 28, 245–248. doi:10.1038/nbt.1614

Palsson, B. Ø. (2011). *Systems biology: Simulation of Dynamic Network States.* Cambridge: Cambridge University Press.

Resasco, D. C., Gao, F., Morgan, F., Novak, I. L., Schaff, J. C., and Slepchenko, B. M. (2012). Virtual cell: computational tools for modeling in cell biology. *Wiley Interdiscip. Rev. Syst. Biol. Med.* 4, 129–140. doi:10.1002/wsbm.165

Rolfsson, O., Palsson, B. Ø., and Thiele, I. (2011). The human metabolic reconstruction recon 1 directs hypotheses of novel human metabolic functions. *BMC Syst. Biol.* 5:155. doi:10.1186/1752-0509-5-155

Sauro, H. M., Hucka, M., Finney, A., Wellock, C., Bolouri, H., Doyle, J., et al. (2003). Next generation simulation tools: the systems biology workbench and BioSPICE integration. *OMICS* 7, 355–372. doi:10.1089/153623103322637670

Sauro, H. M., Karlsson, T. T., Swat, M., Galdzicki, M., and Somogyi, A. (2013). libRoadRunner: a high performance SBML compliant simulator. *Cold Spring Harbor Laboratory.* doi:10.1101/001230

Schaefer, C. F., Anthony, K., Krupa, S., Buchoff, J., Day, M., Hannay, T., et al. (2009). PID: the pathway interaction database. *Nucleic Acids Res.* 37, D674–D679. doi:10.1093/nar/gkn653

Schaff, J. C., Lakshminarayana, A., and Smith, L. P. (2013). "Spatial processes," in *Technical Report.* Pasadena, CA: Caltech. Available from: http://sbml.org/Documents/Specifications/SBML_Level_3/Packages/spatial

Schellenberger, J., Park, J. O., Conrad, T. M., and Palsson, B. Ø. (2010). Bigg: a biochemical genetic and genomic knowledgebase of large scale metabolic reconstructions. *BMC Bioinformatics* 11:213. doi:10.1186/1471-2105-11-213

Schellenberger, J., Que, R., Fleming, R. M. T., Thiele, I., Orth, J. D., Feist, A. M., et al. (2011). Quantitative prediction of cellular metabolism with constraint-based models: the cobra toolbox v2.0. *Nat. Protoc.* 6, 1290–1307. doi:10.1038/nprot.2011.308

Schilstra, M. J., Li, L., Matthews, J., Finney, A., Hucka, M., and Le Novère, N. (2006). CellML2SBML: conversion of CellML into SBML. *Bioinformatics* 22, 1018–1020. doi:10.1093/bioinformatics/btl047

Schmidt, H. (2007). SBaddon: high performance simulation for the systems biology toolbox for MATLAB. *Bioinformatics* 23, 646–647. doi:10.1093/bioinformatics/btl668

Schmidt, H., and Jirstrand, M. (2006). Systems biology toolbox for MATLAB: a computational platform for research in systems biology. *Bioinformatics* 22, 514–515. doi:10.1093/bioinformatics/bti799

Schulz, M., Krause, F., Le Novère, N., Klipp, E., and Liebermeister, W. (2011). Retrieval, alignment, and clustering of computational models based on semantic annotations. *Mol. Syst. Biol.* 7, 512. doi:10.1038/msb.2011.41

Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., et al. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* 13, 2498–2504. doi:10.1101/gr.1239303

Shapiro, B. E., Finney, A., Hucka, M., Bornstein, B. J., Funahashi, A., Jouraku, A., et al. (2007). *Introduction to Systems Biology*, Chapter SBML Models and MathSBML. Totowa, NJ: Humana Press, 395–421.

Shapiro, B. E., Hucka, M., Finney, A., and Doyle, J. (2004). MathSBML: a package for manipulating SBML-based biological models. *Bioinformatics* 20, 2829–2831. doi:10.1093/bioinformatics/bth271

Shen, S. Y., Bergmann, F. T., and Sauro, H. M. (2010). SBML2Ti*k*Z: supporting the SBML render extension in $\LaTeX$. *Bioinformatics* 26, 2794–2795. doi:10.1093/bioinformatics/btq512

Shepherd, G. M., Mirsky, J. S., Healy, M. D., Singer, M. S., Skoufos, E., Hines, M. S., et al. (1998). The human brain project: neuroinformatics tools for integrating, searching and modeling multidisciplinary neuroscience data. *Trends Neurosci.* 21, 460–468. doi:10.1016/S0166-2236(98)01300-9

Smith, L. P., Bergmann, F. T., Chandran, D., and Sauro, H. M. (2009). Antimony: a modular model definition language. *Bioinformatics* 25, 2452–2454. doi:10.1093/bioinformatics/btp401

Smith, L. P., Butterworth, E., Bassingthwaighte, J. B., and Sauro, H. M. (2013a). SBML and CellML translation in antimony and JSim. *Bioinformatics* 30, 903–907. doi:10.1093/bioinformatics/btt641

Smith, L. P., Hucka, M., Hoops, S., Finney, A., Ginkel, M., Myers, C. J., et al. (2013b). "Hierarchical model composition," in *Technical Report*. Available from: http://sbml.org/Documents/Specifications/SBML_Level_3/Packages/comp

Snoep, J. L., and Olivier, B. G. (2003). JWS online cellular systems modelling and microbiology. *Microbiology* 149, 3045–3047. doi:10.1099/mic.0.C0124-0

Sonnenschein, N., and Palsson, B. Ø. (2013). *MASS Toolbox*. Available at: http://opencobra.github.io/MASS-Toolbox/

Swainston, N., Smallbone, K., Mendes, P., Kell, D., and Paton, N. (2011). The SuBliMinaL toolbox: automating steps in the reconstruction of metabolic networks. *J. Integr. Bioinform.* 8, 186. doi:10.2390/biecoll-jib-2011-186

Takizawa, H., Nakamura, K., Tabira, A., Chikahara, Y., Matsui, T., Hiroi, N., et al. (2013). LibSBMLSim: a reference implementation of fully functional SBML simulator. *Bioinformatics* 29, 1474–1476. doi:10.1093/bioinformatics/btt157

Taylor, C. F., Field, D., Sansone, S.-A., Aerts, J., Apweiler, R., Ashburner, M., et al. (2008). Promoting coherent minimum reporting guidelines for biological and biomedical investigations: the MIBBI project. *Nat. Biotechnol.* 26, 889–896. doi:10.1038/nbt.1411

Terfve, C., Cokelaer, T., Henriques, D., MacNamara, A., Goncalves, 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

Thiele, I., Swainston, N., Fleming, R. M. T., Hoppe, A., Sahoo, S., Aurich, M. K., et al. (2013). A community-driven global reconstruction of human metabolism. *Nat. Biotechnol.* 31, 419–425. doi:10.1038/nbt.2488

Töpfer, S., Guthke, R., Driesch, D., Wötzel, D., and Pfaff, M. (2007). "The Net*Gen*erator algorithm: reconstruction of gene regulatory networks," in *Knowledge Discovery and Emergent Complexity in Bioinformatics*, Volume 4366 of *Lecture Notes in Computer Science*, eds K. Tuyls, R. Westra, Y. Saeys, and A. Nowé (Berlin, Heidelberg: Springer).

van Iersel, M. P., Villéger, A. C., Czauderna, T., Boyd, S. E., Bergmann, F. T., Luna, A., et al. (2012). Software support for SBGN maps: SBGN-ML and LibSBGN. *Bioinformatics* 28, 2016–2021. doi:10.1093/bioinformatics/bts270

Vella, M., Cannon, R. C., Crook, S., Davison, A. P., Ganapathy, G., Robinson, H. P. C., et al. (2014). libNeuroML and PyLEMS: using Python to combine procedural and declarative modeling approaches in computational neuroscience. *Front. Neuroinform.* 8:38. doi:10.3389/fninf.2014.00038

Vlaic, S., Hoffmann, B., Kupfer, P., Weber, M., and Dräger, A. (2013). GRN2SBML: automated encoding and annotation of inferred gene regulatory networks complying with SBML. *Bioinformatics* 29, 2216–2217. doi:10.1093/bioinformatics/btt370

Waltemath, D., Adams, R., Beard, D. A., Bergmann, F. T., Bhalla, U. S., Britten, R., et al. (2011a). Minimum information about a simulation experiment (MIASE). *PLoS Comput. Biol.* 7:e1001122. doi:10.1371/journal.pcbi.1001122

Waltemath, D., Adams, R., Bergmann, F. T., Hucka, M., Kolpakov, F., Miller, A. K., et al. (2011b). 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

Waltemath, D., Bergmann, F. T., Chaouiya, C., Czauderna, T., Gleeson, P., Goble, C., et al. (2014). Meeting report from the fourth meeting of the Computational Modeling in Biology Network (COMBINE). *Stand. Genomic Sci.* 9, 1285–1301, doi:10.4056/sigs.5279417

Waltemath, D., Henkel, R., Winter, F., and Wolkenhauer, O. (2013). "Reproducibility of model-based results in systems biology," in *Systems Biology*, eds A. Prokop and B. Csukás (New York: Springer), 301–320.

Wimalaratne, S. M., Halstead, M. D. B., Lloyd, C. M., Cooling, M. T., Crampin, E. J., and Nielsen, P. F. (2009). A method for visualizing CellML models. *Bioinformatics* 25, 3012–3019. doi:10.1093/bioinformatics/btp495

Wittig, U., Rey, M., Kania, R., Bittkowski, M., Shi, L., Golebiewski, M., et al. (2014). Challenges for an enzymatic reaction kinetics database. *FEBS Journal* 281, 572–582. doi:10.1111/febs.12562

Wolstencroft, K., Owen, S., du Preez, F., Krebs, O., Mueller, W., Goble, C., et al. (2011). *The SEEK: A Platform for Sharing Data and Models in Systems Biology*, Vol. 500. San Diego, CA: Elsevier.

Wrzodek, C., Büchel, F., Ruff, M., Dräger, A., and Zell, A. (2013). Precise generation of systems biology models from KEGG pathways. *BMC Syst. Biol.* 7:15. doi:10.1186/1752-0509-7-15

Wrzodek, C., Dräger, A., and Zell, A. (2011). KEGG translator: visualizing and converting the KEGG pathway database to various formats. *Bioinformatics* 27, 2314–2315. doi:10.1093/bioinformatics/btr377

Wrzodek, C., Schröder, A., Dräger, A., Wanke, D., Berendzen, K. W., Kronfeld, M., et al. (2010). Module master: a new tool to decipher transcriptional regulatory networks. *BioSystems* 99, 71–81. doi:10.1016/j.biosystems.2009.09.005

Yu, T., Lloyd, C. M., Nickerson, D. P., Cooling, M. T., Miller, A. K., Garny, A., et al. (2011). The physiome model repository 2. *Bioinformatics* 27, 743–744. doi:10.1093/bioinformatics/btq723

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.

Received: 01 September 2014; accepted: 14 November 2014; published online: 08 December 2014.

Citation: Dräger A and Palsson BØ (2014) Improving collaboration by standardization efforts in systems biology. Front. Bioeng. Biotechnol. ***2****:61. doi: 10.3389/fbioe.2014.00061*

This article was submitted to Systems Biology, a section of the journal Frontiers in Bioengineering and Biotechnology.

ADVANTAGES OF PUBLISHING IN FRONTIERS

FAST PUBLICATION

Average 90 days
from submission
to publication

COLLABORATIVE PEER-REVIEW

Designed to be rigorous –
yet also collaborative, fair and
constructive

RESEARCH NETWORK

Our network
increases readership
for your article

OPEN ACCESS

Articles are free to read,
for greatest visibility

TRANSPARENT

Editors and reviewers
acknowledged by name
on published articles

GLOBAL SPREAD

Six million monthly
page views worldwide

COPYRIGHT TO AUTHORS

No limit to
article distribution
and re-use

IMPACT METRICS

Advanced metrics
track your
article's impact

SUPPORT

By our Swiss-based
editorial team

frontiers

EPFL Innovation Park · Building I · 1015 Lausanne · Switzerland
T +41 21 510 17 00 · info@frontiersin.org · frontiersin.org