ODEbase: A Repository of ODE Systems for Systems Biology

Christoph Lüders
University of Kassel, Germany
chris@cfos.de

Thomas Sturm
CNRS, Inria, and the University of Lorraine, France
MPI Informatics and Saarland University, Germany
thomas.sturm@loria.fr

Ovidiu Radulescu
CNRS and the University of Montpellier, France
ovidiu.radulescu@umontpellier.fr

January 2022

Abstract

Recently, symbolic computation and computer algebra systems have been successfully applied in systems biology, especially in chemical reaction network theory. One advantage of symbolic computation is its potential for qualitative answers to biological questions. Qualitative methods analyze dynamical input systems as formal objects, in contrast to investigating only part of the state space, as is the case with numerical simulation. However, symbolic computation tools and libraries have a different set of requirements for their input data than their numerical counterparts. A common format used in mathematical modeling of biological processes is SBML. We illustrate that the use of SBML data in symbolic computation requires significant pre-processing, incorporating external biological and mathematical expertise. ODEbase provides high quality symbolic computation input data derived from established existing biomodels, covering in particular the BioModels database.

1 Introduction

Recently, symbolic computation methods are playing an increasing role in systems biology and mathematical biology [1, and the references there]. Problems investigated using such methods include Hopf bifurcations, multi-stationarity, multi-scale model reduction, dynamical invariants, and structural properties of

steady state varieties; for details see, e.g., [4, 7, 3, 2, 9, 6]. Compared to numerical analysis and simulation, symbolic computation provides not only quantitative, but qualitative results about network dynamics, to some extent in parametric settings. The biological systems investigated so far had a focus on reaction networks in the sense of chemical reaction network theory [5]. Such networks are usually stored and exchanged in Systems Biology Markup Language (SBML), a free, open, and standardized XML-based format [8].

On the one hand, symbolic computation does not utilize the full information contained in SBML models. For instance, SBML was designed with a focus on network simulation and supports corresponding concepts like events and initial assignments, which are not natural from a formal symbolic computation point of view. On the other hand, symbolic computation operates on formal objects, which are not readily available in SBML. One prominent example are ordinary differential equations (ODEs) describing differential network kinetics along with algebraic constraints, such as conservation laws. The genuine difference between dynamic simulation and static formal analysis requires sensitivity to details and rigor in the course of the construction of symbolic computation input from available SBML descriptions. It is noteworthy that existing SBML parsers generate input for numerical simulation, which is not suited for symbolic computation. MathSBML [16], SBFC [13], SBMLtoODEpy [14], and SBML2Modelica [10] fall into that category.

It is important to understand that the rigorous construction of symbolic computation input poses some substantial problems. Solving, or even recognizing, such problems, requires joint competence and combined efforts not only from biology, but also including mathematics and computer science. We give some examples for such problems:

- SBML allows floating-point values for various entities. However, floating-point values exhibit representation errors and computations are prone to rounding errors. This is inadequate for symbolic computation, where exact computations are performed.

- SBML has liberal naming conventions for species and parameters that interfere with the typically strict rules of symbolic computation software, which are oriented towards mathematical notation. If different users of symbolic computation software rename those identifiers at their own discretion, it becomes cumbersome to compare their results.

- SBML gives modelers versatile opportunities of expression, such as local parameters, function definitions, rules, and initial assignments. For practical reasons, scientific software does not generally support the full SBML feature set. This leads to incompletely imported models or it prohibits the import entirely.

- Symbolic computation is concerned with mathematical properties like deficiency and linear conservation laws, which are available in SBML only

implicitly through computation. Explicit availability is desirable, especially, since some of those computations can become surprisingly time-consuming.

With this in mind, the question is now how to make the abundantly available SBML data accessible for the symbolic computation community in a suitably prepared form.

A natural idea would be to integrate symbolic computation input into the SBML format. However, there are obstacles on both sides: On the symbolic computation side, established software is usually general-purpose, and systems biology is not yet a strong focus of the community. Therefore, widespread support of SBML as an input format for symbolic computation software cannot be expected in the near future. On the systems biology side, the SBML standard would need to be extended. Standardization generally requires considerable efforts, and it seems unlikely that this will be pursued before the links between symbolic computation and systems biology have been further strengthened.

The interdisciplinary project SYMBIONT brings together researchers from mathematics, computer science, and systems biology [1]. Within SYMBIONT, we have started an online database *ODEbase*, which collects symbolic computation input for existing SBML models. All models have been carefully constructed taking into consideration the issues discussed above. Models can be selected by biomathematical properties such as deficiency, rank of the stoichiometric matrix, or numbers of species, parameters, reactions, or conservation constraints. Each model contains a link to its original SBML source. At the time of writing, all our models originate from the BioModels database [12], the world's largest repository of curated mathematical models of biological processes and one of the most important data sources for modeling [11]. Out of the 1044 models from the curated branch of BioModels, we have currently compiled 657 into ODEbase.

As ODEbase has turned out to be extremely valuable throughout the SYMBIONT project, we now make it available to the community as a free and open database, beyond the lifespan of the project.[1] If models require updates, revised versions will be made available, keeping all previous versions for reference. Data can be extracted in Maple, Reduce, SageMath, and LaTeX format. We are open to supporting further formats in the future.

ODEbase provides a canonical source of symbolic computation input related to existing models of biological processes. This has a number of advantages:

1. Interdisciplinary competence: The derivation of adequate ODEs for the kinetics of existing biomodels requires to incorporate external biological and mathematical expertise. We have accomplished this task for a large set of available models and make the results available to the community.

2. Economic use of human resources: Symbolic computation input has been pre-computed and is directly available.

[1] `https://odebase.org`

3. Availability: ODEbase models used and cited in the literature can be conveniently reviewed on the basis of the original data and re-used in follow-up publications.

4. Canonical reference: ODEbase provides an unambiguous mapping of the, in general, too liberal SBML names for species concentrations and parameters to common mathematical notation. This facilitates comparability of results.

5. Benchmarking: Beyond its primary purpose, ODEbase is perfectly suited to generate benchmark sets for novel algorithms and software in the field.

2 Details

2.1 The Content of ODEbase Data Sets

For each model in ODEbase, the following data set is computed from the original SBML input:

Stoichiometric and kinetic matrices. Stoichiometric and kinetic matrices are made explicit. In that course, floating-point values in the SBML input are converted to exact rational numbers.

ODEs for species concentrations. These are explicit first-order, non-linear ODEs that are often, but not necessarily, autonomous. Species are named $x_1, \ldots, x_n$, following common mathematical notation. The ODEs are created from the stoichiometric matrix and the relevant kinetic laws. Again, floating-point values are converted to exact rational numbers. We have taken care to preserve the structure of mathematical terms by using abstract syntax trees as an intermediate format. One visible effect of this are uncanceled rational functions and the presence of stoichiometric coefficients of 1 or -1. If species rate rules are present in the model, the corresponding ODEs are included as well.

Parameter values. Our naming of parameters also follows common mathematical notation, viz., $k_1, \ldots, k_m$. Parameter values are converted from floating-point representation to exact rational numbers. If there are initial assignments or assignment rules in the SBML model, they are applied in the proper order to calculate the parameter values. To avoid any representation errors, all values are queried as text from the XML source.

Map between ODEbase names and original model names. A bijective mapping between the mathematical names for species and parameters and their respective SBML names is provided.

Constraints. All SBML species assignment rules are converted to formal constraints. Furthermore, linear conservation constraints are computed from the stoichiometric matrix using an algorithm by Schuster and Höfer [15], extended to handle multiple model compartments. All constraints introduced this way can be combined with the ODEs mentioned above which yields the relevant ODE system for the model.

Deficiency. The deficiency of the reaction network is computed from its complexes. This is a measure of how independent the reaction vectors are, given the network's linkage class structure [5, Sect. 6.3].

Classification. The ODE right hand sides and both sides of the constraints are analyzed for each model. If all of these terms are polynomials, the model is classified as *polynomial*. If all terms are rational functions, the model is classified as *rational*.

Test for mass-action kinetics. For all models we check whether the SBML-specified kinetics differs from the regular mass-action kinetics [5, Sect. 2.1.2] only by a constant factor. This is a conservative heuristic for identifying models with mass-action kinetics.

2.2 Supported SBML Features

All models in ODEbase are a faithful conversion from the respective SBML model. SBML features recognized during the conservation process include the following:

- species with boundary condition,
- local parameters,
- parameter and species assignment rules,
- parameter and species initial assignments,
- species rate rules,
- function definitions.

SBML supports events, i.e., discrete model changes at certain points in time, and furthermore it supports parameter rate rules. Models that contain either of those are currently not included in ODEbase. Neither are models with irrational parameter values.

Acknowledgements

Andreas Weber, who sadly passed away in 2020, originally gave the idea for ODEbase. Our student assistants Anna Meschede and Matthias Neidhardt at the University of Bonn supported us in the programming the web front-end.

Funding

This work has been supported by the interdisciplinary bilateral project ANR-17-CE40-0036 and DFG-391322026 SYMBIONT [1].

References

[1] François Boulier, François Fages, Ovidiu Radulescu, Satya S. Samal, Andreas Schuppert, Werner M. Seiler, Thomas Sturm, Sebastian Walcher, and Andreas Weber. The SYMBIONT project: Symbolic methods for biological networks. *ACM Commun. Comput. Algebra*, 52(3):67–70, 2018. doi:10.1145/3313880.3313885.

[2] Russell Bradford, James H. Davenport, Matthew England, Hassan Errami, Vladimir Gerdt, Dima Grigoriev, Charles Hoyt, Marek Košta, Ovidiu Radulescu, Thomas Sturm, and Andreas Weber. Identifying the parametric occurrence of multiple steady states for some biological networks. *J. Symb. Comput.*, 98:84–119, 2020. doi:10.1016/j.jsc.2019.07.008.

[3] Alicia Dickenstein. Biochemical reaction networks: An invitation for algebraic geometers. In *Mathematical Congress of the Americas*, volume 656 of *Contemp. Math.*, pages 65–83. 2016. doi:10.1090/conm/656/13076.

[4] Hassan Errami, Markus Eiswirth, Dima Grigoriev, Werner M. Seiler, Thomas Sturm, and Andreas Weber. Detection of Hopf bifurcations in chemical reaction networks using convex coordinates. *J. Comput. Phys.*, 291:279–302, 2015. doi:10.1016/j.jcp.2015.02.050.

[5] Martin Feinberg. *Foundations of Chemical Reaction Network Theory*, volume 202 of *Applied Mathematical Sciences*. Springer, 2019. doi:10.1007/978-3-030-03858-8.

[6] Dima Grigoriev, Alexandru Iosif, Hamid Rahkooy, Thomas Sturm, and Andreas Weber. Efficiently and effectively recognizing toricity of steady state varieties. *Math. Comput. Sci.*, 15(2):199–232, 2021. doi:10.1007/s11786-020-00479-9.

[7] Elizabeth Gross, Heather A. Harrington, Zvi Rosen, and Bernd Sturmfels. Algebraic systems biology: A case study for the Wnt pathway. *Bull. Math. Biol.*, 78(1):21–51, 2016. doi:10.1007/s11538-015-0125-1.

[8] Michael Hucka, Andrew Finney, Herbert M. Sauro, Hamid Bolouri, John C. Doyle, Hiroaki Kitano, Adam P. Arkin, Benjamin J. Bornstein, Dennis Bray, Athel Cornish-Bowden, Autumn A. Cuellar, Serge Dronov, Ernst D. Gilles, Martin Ginkel, Victoria Gor, Igor I. Goryanin, Warren J. Hedley, T. Charles Hodgman, Jan-Hendrik Hofmeyr, Peter J. Hunter, Nick S. Juty, J. L. Kasberger, Andreas Kremling, Ursula Kummer, Nicolas Le Novère,

Leslie M. Loew, Daniel Lucio, Pedro Mendes, Eric Minch, Eric D. Mjolsness, Yasuaki Nakayama, Melanie R. Nelson, Poul F. Nielsen, Takeshi Sakurada, James C. Schaff, Bruce E. Shapiro, Thomas S. Shimizu, Hugh D. Spence, Joerg Stelling, Koichi Takahashi, Masayuki Tomita, John Wagner, and J. Wang. The systems biology markup language (SBML): A medium for representation and exchange of biochemical network models. *Bioinformatics*, 19(4):524–531, 2003. doi:10.1093/bioinformatics/btg015.

[9] Niclas Kruff, Christoph Lüders, Ovidiu Radulescu, Thomas Sturm, and Sebastian Walcher. Algorithmic reduction of biological networks with multiple time scales. *Math. Comput. Sci.*, 15(3):499–534, 2021. doi:10.1007/s11786-021-00515-2.

[10] Filippo Maggioli, Toni Mancini, and Enrico Tronci. SBML2Modelica: Integrating biochemical models within open-standard simulation ecosystems. *Bioinformatics*, 36(7):2165–2172, 2020. doi:10.1093/bioinformatics/btz860.

[11] Rahuman S. Malik-Sheriff, Mihai Glont, Tung V. N. Nguyen, Krishna Tiwari, Matthew G. Roberts, Ashley Xavier, Manh T. Vu, Jinghao Men, Matthieu Maire, Sarubini Kananathan, Emma L. Fairbanks, Johannes P. Meyer, Chinmay Arankalle, Thawfeek M. Varusai, Vincent Knight-Schrijver, Lu Li, Corina Dueñas-Roca, Gaurhari Dass, Sarah M. Keating, Young M. Park, Nicola Buso, Nicolas Rodriguez, Michael Hucka, and Henning Hermjakob. BioModels—15 years of sharing computational models in life science. *Nucleic Acids Res.*, 48(D1):D407–D415, 2019. doi:10.1093/nar/gkz1055.

[12] Nicolas Le Novère, Benjamin J. Bornstein, Alexander Broicher, Mélanie Courtot, Marco Donizelli, Harish Dharuri, Lu Li, Herbert M. Sauro, Maria Schilstra, Bruce E. Shapiro, Jacky L. Snoep, and Michael Hucka. BioModels database: A free, centralized database of curated, published, quantitative kinetic models of biochemical and cellular systems. *Nucleic Acids Res.*, 34(suppl_1):D689–D691, 2006. doi:10.1093/nar/gkj092.

[13] Nicolas Rodriguez, Jean-Baptiste Pettit, Piero Dalle Pezze, Lu Li, Arnaud Henry, Martijn P. van Iersel, Gael Jalowicki, Martina Kutmon, Kedar N. Natarajan, David Tolnay, Melanie I. Stefan, Chris T. Evelo, and Nicolas Le Novère. The systems biology format converter. *BMC Bioinform.*, 17(1):1–7, 2016. doi:10.1186/s12859-016-1000-2.

[14] Steve M. Ruggiero and Ashlee N. Ford Versypt. SBMLtoODEpy: A software program for converting SBML models into ODE models in Python. *J. Open Source Softw.*, 5(53):1643, 2019. doi:10.21105/joss.01643.

[15] Stefan Schuster and Thomas Höfer. Determining all extreme semi-positive conservation relations in chemical reaction systems: A test criterion for conservativity. *J. Chem. Soc., Faraday Trans.*, 87(16):2561–2566, 1991. doi:10.1039/ft9918702561.

[16] Bruce E. Shapiro, Michael Hucka, Andrew Finney, and John C. Doyle. MathSBML: A package for manipulating SBML-based biological models. *Bioinformatics*, 20(16):2829–2831, 2004. doi:10.1093/bioinformatics/bth271.