On the Power of Uncertainties in Microbial System Modeling: No Need To Hide Them Anymore

Benoit Delahaye,[a] Damien Eveillard,[a] Nicholas Bouskill[b]

LS2N, UMR 6004, CNRS, Université de Nantes, ECN, IMTA, Nantes, France[a]; Lawrence Berkeley National Laboratory, Berkeley, California, USA[b]

ABSTRACT For decades, microbiologists have considered uncertainties as an undesired side effect of experimental protocols. As a consequence, standard microbial system modeling strives to hide uncertainties for the sake of deterministic understanding. However, recent studies have highlighted greater experimental variability than expected and emphasized uncertainties not as a weakness but as a necessary feature of complex microbial systems. We therefore advocate that biological uncertainties need to be considered foundational facets that must be incorporated in models. Not only will understanding these uncertainties improve our understanding and identification of microbial traits, it will also provide fundamental insights on microbial systems as a whole. Taking into account uncertainties within microbial models calls for new validation techniques. Formal verification already overcomes this shortcoming by proposing modeling frameworks and validation techniques dedicated to probabilistic models. However, further work remains to extract the full potential of such techniques in the context of microbial models. Herein, we demonstrate how statistical model checking can enhance the development of microbial models by building confidence in the estimation of critical parameters and through improved sensitivity analyses.

KEYWORDS modeling, simulation, uncertainty

Since the work of Monod (1), simple biological modeling has been prominent in microbiology. Because of their experimental tractability and purported simplicity, microbial experimental systems have fostered the rise of several cross-scale modeling approaches from the gene to the population level, which have been extended to test ecoevolutionary hypotheses. These modeling approaches proposed and addressed foundational hypotheses that developed into new biological paradigms such as growth rate or identification of functional units. The first microbial models were driven by reductionist assumptions (e.g., intracellular quota combined with kinetics mimicking biochemistry rules) yet demonstrated remarkable predictive power for portraying the growth of microbes in simple systems such as chemostats (1). Similar quota assumptions were used to model phytoplankton physiology (2) and later for modeling simplified global ocean ecosystems (3).

Reductionist modeling approaches have generally been parameterized from data gleaned from laborious bench experiments. However, contemporary next-generation sequencing (NGS) approaches provide unprecedented characterization of the diversity of microbial communities, yet because feedbacks between biotic and abiotic systems are inherently nonlinear and complex, mathematical models of microbial guilds interacting with their environment are required. Current models have been developed along the lines of systems biology approaches (4, 5) or trait-based models (6). Furthermore, NGS provides a large amount of data that represent one experiment alongside

its associated uncertainties. It is also worth noting that increasing the data set provides a concomitant increase in the number of uncertainties that must be considered.

Within microbial models, these uncertainties can be accounted for by machine learning techniques (see Libbrecht and Noble [7] for a review) that produce automatically predictive (deterministic) models from experimental measurements. Uncertainties are accounted for in the modeling process (e.g., via averaging) but are hidden in the model itself. Moreover, despite the great predictive power of such modelings, the resulting models are not necessarily biologically meaningful. In particular, once parameterized, a model could be overfit to a data set without reflecting emergent properties and precluding or reducing knowledge discovery. Considering that a single microbial system could produce several distinct data sets through several experimental approaches, several different models are built accordingly (8, 9). All corresponding models must then be investigated via automatic learning and verification techniques to take into account their common properties rather than considering each model in isolation (10). Conversely, other probabilistic modelings consider uncertainties but make the parameterization difficult or advocate the use of multiple models that are difficult to validate (11). Despite the aforementioned challenges, uncertainties must be accounted for and integrated into microbial modeling approaches. Such an issue remains a general problem across biology, and even in ecology despite a long tradition of dealing with quantitative uncertainties (12). Herein, as previously done in engineering (13), we advocate that the proper use of dedicated verification techniques could support efforts to capture the complexity of microbial systems within models by promoting a computational convergence on uncertainties rather than simple simulations. Below we present a short overview of current modeling approaches for studying microbial systems, after which we discuss the computational challenges that must be overcome to better take into account uncertainties and verify the resulting models.

MICROBIAL STATE-OF-THE ART MODELINGS AND UNCERTAINTIES

Initial biological modeling efforts were inspired by models of physical systems and formalized using nonlinear ordinary differential equations representing dynamic behaviors of gene activity or molecular concentrations. Interactions within these models are driven by reaction rates associated with particular mechanistic behaviors such as Michaelis-Menten or Hill functions. The need to use formal methods to analyze such models has been discussed extensively in the last 15 years (see De Jong [14] for a state-of-the-art review or Fisher and Henzinger [15]). However, their application mainly results in a discretization of all interactions, such that the model becomes qualitative. An advantage in this context is that the resulting qualitative models are computationally scalable and do not need to incorporate an extensive number of parameters that are mostly out of experimental reach, including parameters that show clear sensitivity to experimental conditions (16). Nonetheless, even if such models are sufficient to represent microbial gene regulatory networks, they are generally not sufficient for modeling quantitative microbial behaviors as needed in the context of simulation of microbial populations and communities, and subsequently, biogeochemical processes.

In order to represent complex quantitative microbial community responses to environmental constraints, several studies postulate that the development of genome-scale metabolic models that consider the whole set of metabolic reactions within a microbial strain are necessary (17). In this context, the metabolic network consists of stoichiometric coefficients and mass balance constraints. Herein, the rates are reduced to fluxes that can be estimated via flux balance analysis (FBA) (18), which negates the need for specific kinetic parameterization yet provides quantitative predictions. FBA models are based on constraints, and solving the metabolic network analysis requires the consideration of boundary conditions for all metabolic fluxes that take place within the cell (see Bordbar et al. [19] for a review). In particular, recently, Basler et al. (20) have shown that those boundaries that reflect uncertainties are necessary to better understand intracellular metabolic flux distribution.

In order to simulate microbial communities, one can consider them at equilibrium

by performing extensions of FBA (21, 22) or focus on dynamic quantitative behaviors by representing kinetics of several microbial populations via traits. These trait-based models use nonlinear differential equations and simplified complex dynamic behaviors with a single parameter. However, because of the inherent complexity of the population (23), estimating such a parameter is a difficult task without considering uncertainties and intensive statistical analyses. Despite several attempts to identify those traits from metagenomic experiments (24), estimating trait parameter values and automatically building predictive models from ecosystem experiments remain key challenges for the field.

NEED FOR DEDICATED MODEL-CHECKING TECHNIQUES

Following standard parameter estimation, whereby simulations satisfy the trends and trajectories of experimental data, models are usually tested through sensitivity analyses. Such analyses test the predictive accuracy of models across a wider range of parameter values. However, while such testing is commonly used in practice because of its computational scalability, sensitivity analysis does not provide a formal guarantee of the correctness of the model but rather a synthesis of extensive number of simulations. In several engineering fields, formal verification of models overcomes this shortcoming by performing model checking (25), which provides such guarantees. Unfortunately, because standard model checking was originally designed to study artificial systems, such as computer programs, it does not necessarily scale up to meet the demands of complexity of microbial systems, in particular when dealing with uncertainties. One must therefore combine model checking with sensitivity analyses, as fostered by recent statistical model-checking (SMC) methods (26–28). Like sensitivity analysis, SMC is based on simulations and executes the model several times to converge on a probability for a given property to be satisfied (see Fig. 1 for an illustration). In this case, the number of simulations that must be performed, and satisfied, is not arbitrarily fixed by the modeler but rather precomputed in order to ensure strong (formal) guarantees on the confidence and error levels of the analysis. Because SMC methods are based on simulations, they do not rely directly on the model structure (i.e., number of variables and constraints), only on the ability to run simulations, regardless of the formalization used (i.e., ordinary differential equations [ODEs], constraint-based, Boolean...). As a consequence, they are suitable for realistic modelings, ensuring strong formal guarantees (29) despite the complexity inherent to microbial systems.

STATISTICAL MODEL CHECKING OF MICROBIAL MODELS

Because uncertainties are central to SMC methods, our hypothesis is that its use will be central for microbial model validation. SMC will indeed build formal confidence (trust) in the validation process while improving standard validation techniques. Standard validation is usually performed using a sufficiently large number of model simulations, but the precise number is left to the acumen of the modeler, which is not a satisfying guarantee with respect to the precision and correctness of the analyses. In contrast, the precision and correctness of SMC methods are formally certified by using statistical results to compute the required number of model simulations. Moreover, SMC methods are tailored for the analysis of models that incorporate uncertainties *per se* and therefore take into account parameter variations as standard characteristics of the models studied. Thus, two potential SMC applications in the context of microbial system validation could be emphasized.

Model certification rather than sensitivity analysis. Instead of fixing parameter values to their mean observed values and performing sensitivity analysis of one parameter at a time (Fig. 1A), we propose to embed the uncertainty of the parameter values into the models by assigning each parameter to a probability distribution based on its potential values informed by lab or field experiments (Fig. 1B). Consequently, a trait must then be considered a distribution on a range of values instead of a single value that represents multiple experiments. The distribution of values over a given
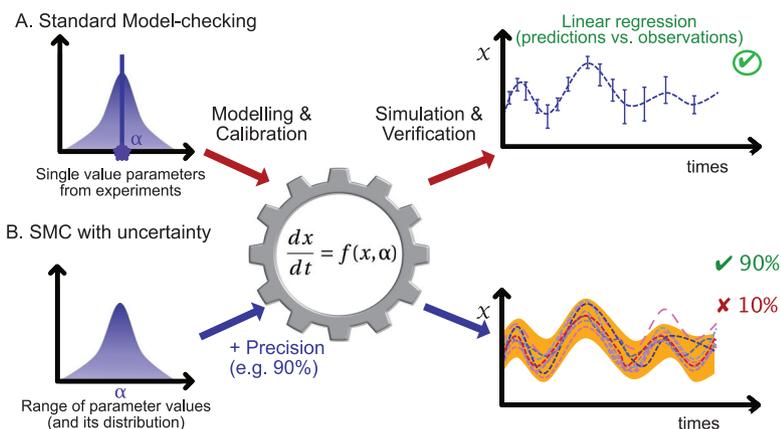
FIG 1 Illustration of model checking without and with uncertainties. Following a range of experiments, standard data analysis highlights the distribution of values for a given parameter α. (A) Along a standard model-checking protocol, one assumes a single value for α, usually the mean. Such a value is then used for model calibration, which allows a simulation. Simulation results are then compared with observations for the sake of model verification (e.g., usually via linear regression between prediction and observations). (B) An example of a model-checking protocol that considers uncertainties *per se*. Instead of considering a single parameter value, one considers a range of values and precision guarantees and performs a range of simulations accordingly (one per color). Altogether, this SMC approach validates the models while taking into account intrinsic uncertainties and guarantees the desired precision (90% here).

range can thus be set as Gaussian, or as another distribution, in order to fit experimental uncertainties. In SMC, model simulations can be performed by picking parameter values within their attached distributions (i.e., by considering the variances of all parameters) and executing standard simulations. Following a simulation, deciding whether one prediction is valid can be done by computing the deviation of predicted values from existing observations. Since several simulations are performed (for several parameter values), SMC outputs a score representing the ratio of valid simulations. Thus, the SMC method performs a generalization of standard sensitivity analyses, not by analyzing the sensitivity of a single average simulation but by analyzing all feasible simulations and proposing general statistics of the whole; i.e., accurate statistical guarantees to perform predictive simulations while taking into account experimental uncertainties. By extension, considering uncertainties also allows us to certify all simulation behaviors (i.e., average of simulations) for the sake of model validation, rather than validating a single behavior (i.e., simulation of an average).

Model parameter estimation with uncertainties. Model parameters are often difficult to measure because microbial communities encapsulate an array of trait data leading to wide ranges in parameter values. Moreover, identifying a distribution of parameter values, for instance, trait parameter values for models of microbe-mediated biogeochemical processes remains a difficult task. One could overcome this difficulty by considering a slight modification of the SMC paradigm to decipher the global set of parameters (with uncertainties) that best fit the experimental data (Fig. 2). While the global range of potential parameter values is often known, the aim here is to identify, within this range, the "ideal" parameter values (i.e., those that produce the highest rate of valid simulations). In this context, a partition of the global range of parameter values (alongside the variance) is produced (Fig. 2B), and simulations are performed for each of the obtained "subranges" (Fig. 2C). SMC can then provide a certified score for each of those subranges, which represent an evaluation of their adequacy with respect to experimental data (Fig. 2D and E). When performed on all subranges from the partition, this method allows us to identify the parameter subranges that best fit the experimental data.

CONCLUSION

Uncertainties in parameter values are inherent to models built from experimental data. Instead of doing our best to rule out these uncertainties through deterministic
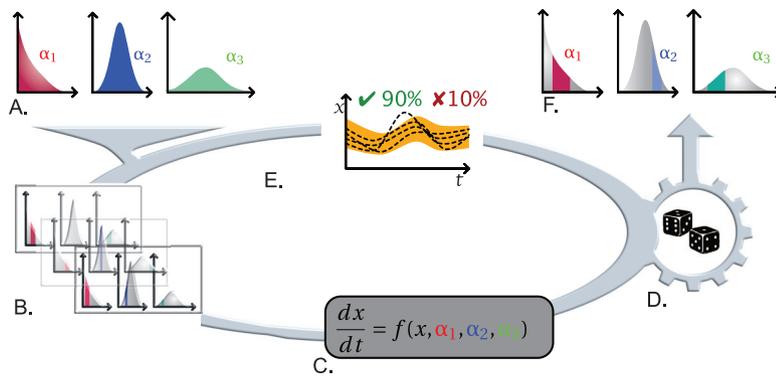
FIG 2 Statistical model checking (SMC) for model parameter estimation with uncertainties. (A and B) Considering the distribution of parameter values (A), SMC will perform a partition of the global range of parameter values (B). Notably, all parameters will be identified as a whole, instead of identifying each parameter independently from others as in standard parameterization techniques. (C and D) Probabilistic simulations are then performed for each of the "subranges" obtained. (E) Simulations of all models are then compared to experimental data for the sake of adequacy estimation. (F) Iterated several times, this protocol allows us to identify parameter subranges that, considered altogether, best fit the experimental data.

modeling, we advocate that they should be incorporated into the models via the development of dedicated probabilistic modelings. Originally developed for software applications, statistical verification of such models will enhance the accuracy of model validation while also bringing formal evidence of the correctness of the approach. In addition, the future development of dedicated SMC methods will be the necessary steps to estimate parameter values with uncertainties, ensuring the satisfaction of desired properties, which represents the next methodological block for modeling quantitatively microbial systems from genes to ecosystems.

ACKNOWLEDGMENT

REFERENCES

1. Monod J. 1949. The growth of bacterial cultures. Annu Rev Microbiol 3:371–394. https://doi.org/10.1146/annurev.mi.03.100149.002103.
2. Droop M. 1973. Some thoughts on nutrient limitation in algae. J Phycol 9:264–272.
3. Follows MJ, Dutkiewicz S. 2011. Modeling diverse communities of marine microbes. Annu Rev Mar Sci 3:427–451. https://doi.org/10.1146/annurev-marine-120709-142848.
4. Raes J, Bork P. 2008. Molecular eco-systems biology: towards an understanding of community function. Nat Rev Microbiol 6:693–699. https://doi.org/10.1038/nrmicro1935.
5. Thiele I, Heinken A, Fleming RMT. 2013. A systems biology approach to studying the role of microbes in human health. Curr Opin Biotechnol 24:4–12. https://doi.org/10.1016/j.copbio.2012.10.001.
6. Litchman E, Klausmeier CA. 2008. Trait-based community ecology of phytoplankton. Annu Rev Ecol Evol Syst 39:615–639. https://doi.org/10.1146/annurev.ecolsys.39.110707.173549.
7. Libbrecht MW, Noble WS. 2015. Machine learning in genetics and genomics. Nat Rev Genet 16:321–332. https://doi.org/10.1038/nrg3920.
8. Simão E, Remy E, Thieffry D, Chaouiya C. 2005. Qualitative modelling of regulated metabolic pathways: application to the tryptophan biosynthesis in E. coli. Bioinformatics 21(Suppl 2):ii190–ii196. https://doi.org/10.1093/bioinformatics/bti1130.
9. Salazar-Cavazos E, Santillán M. 2014. Optimal performance of the tryptophan operon of E. coli: a stochastic, dynamical, mathematical-modeling approach. Bull Math Biol 76:314–334. https://doi.org/10.1007/s11538-013-9920-8.
10. Vert J-P. 2010. Reconstruction of biological networks by supervised machine learning approaches, p 165–188. In Lodhi HM, Muggleton SH (ed), Elements of computational systems biology. John Wiley & Sons, Inc, Hoboken, NY.
11. Friedman N, Linial M, Nachman I, Pe'er D. 2000. Using Bayesian networks to analyze expression data. J Comput Biol 7:601–620. https://doi.org/10.1089/106652700750050961.
12. Mouquet N, Lagadeuc Y, Devictor V, Doyen L, Duputié A, Eveillard D, Faure D, Garnier E, Gimenez O, Huneman P, Jabot F, Jarne P, Joly D, Julliard R, Kéfi S, Kergoat GJ, Lavorel S, Le Gall L, Meslin L, Morand S, Morin X, Morlon H, Pinay G, Pradel R, Schurr FM, Thuiller W, Loreau M. 2015. Predictive ecology in a changing world. J Appl Ecol 52:1293–1310. https://doi.org/10.1111/1365-2664.12482.
13. Zuliani P, Platzer A, Clarke EM. 2013. Bayesian statistical model checking with application to Stateflow/Simulink verification. Formal Methods Syst Des 43:338–367. https://doi.org/10.1007/s10703-013-0195-3.
14. De Jong H. 2002. Modeling and simulation of genetic regulatory systems: a literature review. J Comput Biol 9:67–103. https://doi.org/10.1089/10665270252833208.
15. Fisher J, Henzinger TA. 2007. Executable cell biology. Nat Biotechnol 25:1239–1249. https://doi.org/10.1038/nbt1356.
16. Edwards JS, Palsson BO. 2000. Robustness analysis of the Escherichia coli metabolic network. Biotechnol Prog 16:927–939. https://doi.org/10.1021/bp0000712.
17. Thiele I, Palsson BØ. 2010. A protocol for generating a high-quality genome-scale metabolic reconstruction. Nat Protoc 5:93–121. https://doi.org/10.1038/nprot.2009.203.
18. Orth JD, Thiele I, Palsson BØ. 2010. What is flux balance analysis? Nat Biotechnol 28:245–248. https://doi.org/10.1038/nbt.1614.
19. Bordbar A, Monk JM, King ZA, Palsson BO. 2014. Constraint-based

mSystems®

models predict metabolic and associated cellular functions. Nat Rev Genet 15:107–120. https://doi.org/10.1038/nrg3643.

20. Basler G, Nikoloski Z, Larhlimi A, Barabási AL, Liu YY. 2016. Control of fluxes in metabolic networks. Genome Res 26:956–968. https://doi.org/10.1101/gr.202648.115.

21. Biggs MB, Papin JA. 2017. Managing uncertainty in metabolic network structure and improving predictions using Ensemble FBA. PLoS Comput Biol 13:e1005413. https://doi.org/10.1371/journal.pcbi.1005413.

22. Budinich M, Bourdon J, Larhlimi A, Eveillard D. 2017. A multi-objective constraint-based approach for modeling genome-scale microbial ecosystems. PLoS One 12:e0171744. https://doi.org/10.1371/journal.pone.0171744.

23. Loreau M, Naeem S, Inchausti P, Bengtsson J, Grime JP, Hector A, Hooper DU, Huston MA, Raffaelli D, Schmid B, Tilman D, Wardle DA. 2001. Biodiversity and ecosystem functioning: current knowledge and future challenges. Science 294:804–808. https://doi.org/10.1126/science.1064088.

24. Raes J, Letunic I, Yamada T, Jensen LJ, Bork P. 2011. Toward molecular trait-based ecology through integration of biogeochemical, geographical and metagenomic data. Mol Syst Biol 7:473. https://doi.org/10.1038/msb.2011.6.

25. Clarke EM, Jr, Grumberg O, Peled DA. 1999. Model checking. MIT Press, Cambridge, MA.

26. Legay A, Delahaye B, Bensalem S. 2010. Statistical model checking: an overview, p 122–135. In Barringer H et al. (ed), Runtime verification. RV 2010. Lecture Notes in Computer Science, vol 6418. Springer, Berlin, Germany. https://doi.org/10.1007/978-3-642-16612-9_11.

27. Koh CH, Palaniappan SK, Thiagarajan PS, Wong L. 2012. Improved statistical model checking methods for pathway analysis. BMC Bioinformatics 13:S15. https://doi.org/10.1186/1471-2105-13-S17-S15.

28. Younes HLS, Simmons RG. 2006. Statistical probabilistic model checking with a focus on time-bounded properties. Inform Comput 204:1368–1409. https://doi.org/10.1016/j.ic.2006.05.002.

29. Kwiatkowska M, Norman G, Parker D. 2008. Using probabilistic model checking in systems biology. ACM SIGMETRICS Perform Eval Rev 35:14–21.