De novo Inference of Stochastic Mechanisms

I. Chattopadhyay *, A. Kuchina †, G. Suel † and H. Lipson *‡

***Short Abstract*—Gillespie simulation is used extensively to investigate stochastic phenomena. The inverse problem, however, has remained largely unsolved: How to reconstruct underlying reactions *de novo* from sparse, intermittent and aggregate observations. We argue that, under specific assumptions, the set of relative population updates in phase space forms a convex polytope whose vertices are indicative of the dominant reactions. Using this simple principle we reconstruct stochastic reaction structure and propensities from a variety of simulated and experimental systems, where possibly thousands of reactions may be occurring in between measurement of aggregate population counts. In addition to *de novo* model structure inference, this technique is shown to yield key mechanistic insights into complex regulatory circuits, $e.g.$, the ones governing competence dynamics of B. subtilis, and the predator-prey dynamics of Paramecium-Didinum experimental ecosystem.**

***Keywords*—Stochastic Simulation, *De novo* Inference**

Detailed Abstract

Gillespie's stochastic simulation algorithm (SSA) [1] is an accurate procedure for predicting time-evolution of finite interacting populations of chemical species in continuous time; essentially defining a computational approach to indirectly solve the intractable probabilistic master equation in the forward direction. SSA along with its generalizations [2] form the basis of most stochastic kinetic modeling techniques [3]. Inspite of providing an elegant technique to simulate a completely specified Markov Jump Process (MJP) [4], as is required to simulate chemical interaction of discrete agents, it is not particularly clear how observed time series may be incorporated in a principled manner in Gillespie's SSA; or how, if at all, the procedure may be reversed to distill the likely reaction structure from such observed data.

In the trivial case where individual reactions are directly observable, the most likely set of reactions can be determined easily. However, population changes are usually only observed intermittently. Gillespie's formulation indicates that the probabilities of individual reactions are possibly nonlinear functions of instantaneous population numbers; and since transpired reactions change the population counts of the participating species, each successive occurrence of the same reaction transpires with a possibly different occurrence probability; making identification notoriously difficult. Despite significant headway into the *calibration problem*, $i.e.$ estimation of parameters, given a model structure and observed expression data [5], [6], [7], *de novo* inference of reaction structure has proved to be elusive.

The present work delineates a new principle to reverse-engineer observed population time series for *de novo* structural identification along with estimation of reaction propensities (See Fig. 1). We only need intermittent measurements

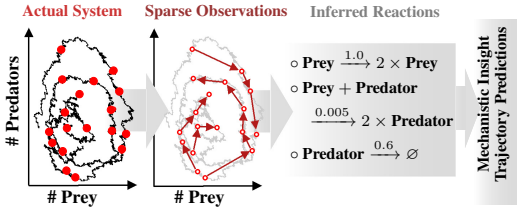

Fig. 1. **Reverse engineering reaction model from observable 'way-points'** (Left) Aggregate population counts are observed intermittently (red circle way-points). From these dots alone (Middle), we reconstruct the most likely generative process (Right) underlying the data, leading to mechanistic insight. Actual data and automatically-inferred system shown.

of the system state represented as population counts of each participating species, and may skip many (in the order of thousands in some of our examples) reactions between successive measurements. For a system evolving stochastically, with the population counts of the interacting chemical species or agents observed intermittently, we rigorously establish that the set of relative update vectors approximately define a convex bounded polytope. The direction cosine of each vertex of this polytope coincides with the population update realized by a major driving reaction. Furthermore, this alignment is independent, upto a certain point, of the degree of intermittency of the observations. We show that these polytope vertices, and hence the direction cosines of the hidden dominant reactions, may be identified from the observed population time series alone, with no a priori system knowledge; thus defining a novel approach for *de novo* structure identification.

We assume a well-mixed system with constant propensities. Additionally, the system must admit a probabilistic equilibrium, and even then, only the dominant reactions are identifiable. Even with these limitations, the present work is, to the best of our knowledge, the only *de novo* structural inference technique for stochastic evolution in the literature requiring no apriori assumptions on the model structure; and is demonstrated on both simulated data, and on experimental ecological and biochemical systems (Paramecium-Didinum predator-prey dynamics, and interaction of key regulators in the competence dynamics of *B. subtilis*).

References

[1] D. T. Gillespie, "Exact stochastic simulation of coupled chemical reactions," *The Journal of Physical Chemistry*, vol. 81, no. 25, pp. 2340–2361, 1977.

[2] ——, "Approximate accelerated stochastic simulation of chemically reacting systems," *J Chem Phys*, vol. 115, pp. 1716–1733, 2001.

[3] D. Wilkinson, *Stochastic Models for System Biology*, ser. Chapman and Hall/CRC Mathematical and Computational Biology Series.

[4] M. Opper and G. Sanguinetti, "Variational inference for markov jump processes in nips 2007," in *NIPS*, J. C. Platt, D. Koller, Y. Singer, and S. T. Roweis, Eds. Curran Associates, Inc., 2007.

[5] R. J. Boys, D. J. Wilkinson, and T. B. Kirkwood, "Bayesian inference for a discretely observed stochastic kinetic model," *Statistics and Computing*, vol. 18, no. 2, pp. 125–135, Jun. 2008.

[6] A. Golightly and D. J. Wilkinson, "Bayesian inference for nonlinear multivariate diffusion models observed with error," *Comput. Stat. Data Anal.*, vol. 52, no. 3, pp. 1674–1693, Jan. 2008.

[7] P. Milner, C. S. Gillespie, and D. J. Wilkinson, "Moment closure based parameter inference of stochastic kinetic models," *Statistics and Computing*, pp. 1–9, 2012.

* Department of Mecchanical & Aerospace Engg., Cornell University
† Section Of Molecular Biology, University of California, San Diego
‡ Computing & Information Science, Cornell University

Supported in part by the US Army Research Office Biomathematics program (#W911NF-12-1-0499).