Published in final edited form as:

J Phys Chem B. 2018 June 21; 122(24): 6351–6356. doi:10.1021/acs.jpcc.8b02960.

Using Equation-Free Computation to Accelerate Network-Free Stochastic Simulation of Chemical Kinetics

Yen Ting Lin[†], Lily A. Chylek^{†,‡}, Nathan W. Lemons[†], and William S. Hlavacek^{†,*}

[†]Theoretical Division and Center for Nonlinear Studies, Los Alamos National Laboratory, New Mexico 87545, USA

Abstract

The chemical kinetics of many complex systems can be concisely represented by reaction rules, which can be used to generate reaction events via a kinetic Monte Carlo method that has been termed network-free simulation. Here, we demonstrate accelerated network-free simulation through a novel approach to equation-free computation. In this process, variables are introduced that approximately capture system state. Derivatives of these variables are estimated using short bursts of exact stochastic simulation and finite differencing. The variables are then projected forward in time via a numerical integration scheme, after which a new exact stochastic simulation is initialized and the whole process repeats. The projection step increases efficiency by bypassing the firing of numerous individual reaction events. As we show, the projected variables may be defined as populations of building blocks of chemical species. The maximal number of connected molecules included in these building blocks determines the degree of approximation. Equation-free acceleration of network-free simulation is found to be both accurate and efficient.

Graphical Abstract

wish@lanl.gov, Phone: +1 505-665-1355.

[‡]Current affiliation: Department of Systems Biology, Harvard Medical School, Boston, MA 02115, USA

Introduction

Many processes in chemistry and biology have the potential to rearrange molecular building blocks in a combinatorial number of ways. These processes include combinatorial organic synthesis, hydrocarbon pyrolysis, oxidation/combustion, metabolism, and cell signaling.¹ When studying systems marked by combinatorial complexity, it can be convenient to characterize state transitions in terms of reaction rules,² or generalized reactions. A reaction rule defines a transformation (e.g., a functional group transformation) and the necessary and sufficient properties of reactants. This information is used to identify the chemical species to which a rule can be applied and to determine the products that result from applying the transformation of the rule. The advantage of this approach is that a rule can be defined without needing to fully identify the particular set of chemical species that participate in a transformation, which allows for compact model formulation. Often, a rule is written such that numerous chemical species qualify as reactants. Rules have been used to analyze and reason about diverse (bio)chemical systems with the aid of various theoretical frameworks and software tools.^{1–3}

In the simplest use cases, rules enable simulations of well-mixed chemical kinetics, which is our focus here. For example, starting from a set of specified seed species, rules may be applied to these species to automatically generate the reactions implied. Repeated applications of rules to new products will yield a chemical reaction network. If rules are associated with rate laws, and rule-derived reactions are taken to inherit these rate laws, derived reactions can be used as generators of reaction events in a stochastic simulation of discrete-event chemical kinetics or, alternatively, ordinary differential equations (ODEs) can be formulated and then numerically integrated.^{2,4} In another approach, rules themselves can serve as generators of reaction events in a stochastic simulation, because rules can be assigned rates if the configuration of a system is tracked.⁵ This latter type of approach, which has been called network-free simulation (because network generation is not required), becomes necessary or attractive when rules imply a sufficiently large reaction network, as when rules define polymerization reactions.^{2,5} A drawback of network-free simulation is that this approach becomes inefficient, possibly prohibitively so, if the number of events per unit time is large, as when chemical species have large population sizes. This is because system state is advanced only one reaction event at a time. Thus, to make network-free simulation applicable to a larger array of problems, there is a need for methods that improve its efficiency.

Methods

Here, we report algorithms that accelerate network-free stochastic (i.e., kinetic Monte Carlo) simulation of rule-defined, well-mixed chemical kinetics by leveraging the concept of equation-free computation.⁶ Equation-free computation, illustrated in Figure 1, is based on the observation that numerical integration does not require any knowledge of the equations being integrated if derivative information for the equations' variables can be obtained through other means. In modeling of chemical kinetics, derivative information can be obtained through finite difference approximations based on data generated in short bursts of exact stochastic simulation. The derivative information can be used to project the system

state forward in time. This approach is particularly useful for accelerating simulation if one can identify projected variables that provide a coarse but sufficiently accurate representation of system state, such that not all details of the system state need to be considered during forward projection.

The passing of information from projected variables to a system configuration is called lifting, and the passing of information from a system configuration to projected variables is called restriction. Restriction is straightforward because it consists of calculating properties of a known system state. Finding suitable projected variables, and suitable ways to perform lifting, is often the crux of making equation-free computation useful in a specific application. In applications, coarse projected variables are typically found by detecting and taking advantage of time-scale separation.⁶ Here, we propose a new approach to this problem.

We define coarse projected variables as populations of connected sets of the molecular building blocks of chemical species, i.e., populations of chemical moieties. Identification of projected variables begins with defining the possible states of individual molecules. Moiety variables can be systematically expanded by considering the states of larger assemblies of molecules. For example, first-degree variables track the possible states of isolated molecules, second-degree variables track the states of all possible connected pairs of molecules, etc.

A system is configured for simulation by assuming that the moieties corresponding to projected variables are independently distributed, i.e., that the population of each moiety is unaffected by the populations of other moieties.

The choice of the set of moieties is not unique, and the choice entails a trade-off between efficiency and accuracy. When the set is small (e.g., when only one molecule is included in each moiety) the correlations introduced by contextual constraints on interactions (e.g., how the bound state of one molecule affects its interaction with another) will not be captured. Thus, the populations of moieties will not truly be independent of each other as assumed. On the other hand, when the set is large, more contextual constraints will be captured, but the computational cost of identifying the moieties following exact simulation will reduce, and potentially compromise, computational efficiency. In rule-based modeling, interactions between molecules are commonly described in a modular sense (i.e., such that interactions depend on context in a limited way), so once a sufficiently large set of moieties is adopted, further expansion of the set will not improve accuracy.

To make these ideas more concrete, let us consider a simple model for assembly of a linear aggregate (Figure 2). Figure 2(a) provides a schematic diagram of the model which consists of six types of indivisible molecules: *A*, *B*, *C*, *D*, *E*, and *F*. Each molecule can bind to others via specific binding sites, and these sites are named according to which molecule they bind to (e.g., site *b* in molecule *A* binds to molecule *B*). There are a total of 21 possible species.

The minimal set of coarse-grained patterns, the first-degree moieties, comprise single molecules and all their possible states. For example, there are four first-degree moieties that are associated with molecule *B*: *B* bound at site *a* but not at site *c*, *B* bound at *c* but not *a*, *B* bound at both sites, and *B* bound at neither site. These moieties are illustrated in Figure 2(b).

Second-degree moieties comprise two-molecule assemblies with all possible states of the constituent molecules. Two examples of second-degree moieties are shown in Figure 2(c). Both of these moieties consist of B bound to A , but differ in whether B is bound to C .

In some cases, the population of a moiety may be derived from the population of another. For example, because the model of Figure 2(a) does not involve import, export, synthesis, or degradation of molecules, the total populations of all molecules are conserved. Therefore, we can express the population of the first moiety in Figure 2(b) as the total quantity of molecule B minus the quantities of the other three B -associated moieties. As a result, the first moiety can (optionally) be removed from the set of projected variables without losing information.

Once moiety variables have been calculated from the configuration of a system and new values are obtained through projection forward in time (e.g., via Eulers method), the new values of the moiety variables are used in lifting. Lifting is the process by which the system's detailed configuration, i.e., the populations of all chemical species, is computed from the projected variables. Some species correspond directly to projected variables. The populations of other, larger complexes can be computed if their populations can be expressed in terms of independently distributed constituent complexes. For example, consider a three-molecule complex consisting of A bound to B bound to C . Note that this complex excludes molecule D . This complex's population is not directly represented by any primary or secondary variable, but can still be computed from them under an assumption of independence. Let the populations of B bound to A and C , C bound to B and D , and C bound only to B , be represented with $N_{a,c}^B$, $N_{b,d}^C$, and N_b^C , respectively. The population of A bound to B bound to C is approximated by

$$N^{ABC} \approx N_{a,c}^B \frac{N_b^C}{N_{b,d}^C + N_b^C}, \quad (1)$$

which is derived by assuming that the probability of a C molecule *not* binding to a D is independent of whether a C molecule is bound to a B molecule.

Useful relationships exist among the variables. For example, the total amount of B bound to C is equal to the total amount of C bound to B . Mathematically, $N_c^B + N_{a,c}^B = N_{b,d}^C + N_b^C$.

Using this relationship, we can derive the following alternative expression for the population of N^{ABC} :

$$N^{ABC} \approx N_b^C \frac{N_{a,c}^B}{N_c^B + N_{a,c}^B}. \quad (2)$$

The population of any larger complex is also approximated on the basis of independence assumptions. For example, the largest assembly in the model of Figure 2, which is a six-molecule complex, has an approximate population given by:

$$N^{ABCDEF} \approx N_{a,c}^B \frac{N_{b,d}^C}{N_{b,d}^C + N_b^C} \frac{N_{c,e}^D}{N_{c,e}^D + N_c^D} \frac{N_{d,f}^E}{N_{d,f}^E + N_d^E}, \quad (3)$$

where N_c^D , $N_{c,e}^D$, N_d^E , and $N_{d,f}^E$ are the populations of D bound only to C , D bound to C and E , E bound only to D , and E bound to D and F , respectively.

The above idea can be generalized to a set of K^{th} -degree moiety variables, which give the populations of connected sets of building blocks consisting of 1 to K molecules (Figure 2(a)). Higher-degree moieties may be introduced to capture correlations introduced by the reaction rules of a model. The populations of larger moieties (containing a number of molecules greater than K) are, as before, approximated by assuming independence. It should be pointed out that the set of K^{th} -degree moieties must account for connected sets of 1, 2, ... K molecules because, for example, if only the populations of all the three-molecule moieties are provided, the population of a dimer is not defined. When K is equal to the number of molecules in the largest possible complex, all chemical species in the system are accounted for. Consideration of such a case is only feasible when the full reaction network can be generated.

Results

To demonstrate the methodology described above, we apply it to a published rule-based model for the dynamics of early events in signaling by the epidermal growth factor (EGF) receptor (EGFR).⁸ The model considers five macromolecules (proteins), binding sites responsible for interactions among the macromolecules, and selected sites within the macromolecules that undergo covalent modification (phosphorylation). A contact map showing the interactions between the molecules is presented in Figure 3(a). The rules of the model imply 356 possible chemical species and 3,749 possible unidirectional reactions between these species. The rules impose contextual constraints on interactions, which induces correlations between the states of sites considered in the model. Thus, the model serves as a good illustration of how the K^{th} -degree moiety variables provide more accuracy as K increases. Schematic diagrams of the $K = 2$ degree moieties are presented in Figure 3(c–d). The full network can be generated, and we adopt the deterministic dynamics described by the ODEs generated from model rules by the BioNetGen software package⁹ as the ground truth.

We present the evolution of populations of selected chemical species in Figure 4 with acceleration of stochastic simulation for various values of K . For all the populations we have chosen to observe, the moiety-matching method reasonably approximates the ground truth when $K = 3$. When $K = 4$, the set of moieties is a complete set that can be used to exactly reconstruct the detailed configuration of the system. Therefore, the errors of $K = 4$ simulation are entirely attributable to finite differencing errors.

We compared the efficiency of equation-free methods (using moiety variables of first degree, second degree, etc.) against the efficiency of exact stochastic simulation for the model of Blinov et al.⁸ To consider a range of simulation costs, in comparisons, we scaled system size from 10% of original size to 100-fold larger than original size. Efficiency comparisons are summarized in Figure 5. Because all the species can be enumerated with this example, the efficiency of the exact stochastic simulation is optimized and serves as a lower-bound on efficiency of the network-free simulation.¹¹ As shown in Figure 5, the equation-free methods increase simulation efficiency.

It is peculiar why the fourth-order pattern matching method performs most efficiently among the various orders of pattern matching methods. As mentioned above, the set of moieties with $K = 4$ fully describes the detailed configuration of the system. Hence, there is no computational cost during restriction and lifting because the full list of the species are generated before each short burst of exact simulation. Thus, equation-free computation at this order efficiently obtains a system state equivalent to the state generated through exact simulation. In comparison, lower-order pattern matching methods were less efficient because of the need to perform additional operations to identify the coarse-grained moiety variables. We expect that the efficiency of the lower-order pattern matching methods will be more efficient in cases where generating the full network is not possible.

Discussion and Conclusion

In summary, we present a novel form of equation-free computation to increase the efficiency of event-driven simulations for (large) chemical reaction networks described by reaction rules. Equation-free methods originated in numerical analyses for multiscale models. These methods typically provide a coarse-grained dynamical picture of multiscale models and improve the efficiency of a simulation when projective integration is introduced. We describe a way of identifying chemical moieties whose populations can act as coarse-grained variables. In the lifting operation, the values of these moiety variables are used to reconstruct the detailed configuration of a system, possibly with reliance on assumptions of independence. An independence assumption, if used, is not always valid; consequently, lifting errors may be introduced. Thus, when K is below a threshold, at which all correlations are captured by moiety variables, the methods described here trade accuracy for efficiency.

Errors can be reduced by taking into account additional correlations between molecular states (induced by contextual constraints in rules) through inclusion of higher-degree moieties in the set. The set of moieties can be expanded systematically and tested for convergence in a trial-and-error manner. When the order of the set of moieties is sufficiently large, all correlations will be captured in the set, and the coarse-grained variables can be used to completely generate system configurations without loss of information.

Acknowledgments

LAC, NWL, and YTL were supported by the Center for Nonlinear Studies and the Laboratory-Directed Research and Development program at Los Alamos National Laboratory, which is operated for the National Nuclear Security Administration of the U.S. Department of Energy under contract DE-AC52-06NA25396. YTL and WSH were also

supported by grants from the National Institutes of Health (R01CA197398 and R01GM111510). LAC acknowledges support from the Defense Advanced Research Projects Agency (W911NF-14-1-0397).

References

1. de Oliveira Lus P, Hudebine Damien, Guillaume Denis, Verstraete J. A review of kinetic modeling methodologies for complex processes. *Oil Gas Sci Technol*. Jan.2016 71:45.
2. Chylek LA, Harris LA, Tung CS, Faeder JR, Lopez CF, Hlavacek WS. Rule-based modeling: a computational approach for studying biomolecular site dynamics in cell signaling systems. *Wiley Interdiscip Rev Syst Biol Med*. 2014; 6:13–36. [PubMed: 24123887]
3. Chylek LA, Stites EC, Posner RG, Hlavacek WS. *Systems Biology* Prokop A, Csukás B, editors Vol. Chapter 9. Springer; Dordrecht: 2013 273300
4. Faulon JL, Sault AG. Stochastic generator of chemical structure. 3. Reaction network generation. *J Chem Inf Comput Sci*. 2001; 41:894–908. [PubMed: 11500106]
5. Yang J, Monine MI, Faeder JR, Hlavacek WS. Kinetic Monte Carlo for rule-based modeling of biochemical networks. *Phys Rev E*. 2008; 78:031910.
6. Kevrekidis IG, Samaey G. Equation-free multiscale computation: algorithms and applications. *Annu Rev Phys Chem*. 2009; 60:321–344. [PubMed: 19335220]
7. Faeder JR, Blinov ML, Hlavacek WS. Rule-based modeling of biochemical systems with BioNetGen. *Methods Mol Biol*. 2009; 500:113–167. [PubMed: 19399430]
8. Blinov ML, Faeder JR, Goldstein B, Hlavacek WS. A network model of early events in epidermal growth factor receptor signaling that accounts for combinatorial complexity. *Bio Systems*. 2006; 83:136–151.
9. Harris LA, Hogg JS, Tapia JJ, Sekar JA, Gupta S, Korsunsky I, Arora A, Barua D, Sheehan RP, Faeder JR. BioNetGen 2. 2: advances in rule-based modeling. *Bioinformatics*. 2016; 32:3366–3368. [PubMed: 27402907]
10. Gillespie DT. A general method for numerically simulating the stochastic time evolution of coupled chemical reactions. *J Comput Phys*. 1976; 22:403–434.
11. Suderman R, Mitra ED, Lin YT, Erickson KE, Feng S, Hlavacek WS. Generalizing Gillespie's direct method to enable network-free simulations. *Bull Math Biol*. in press.

Figure 1.

Illustration of the basic concepts of equation-free computation through simulation of stochastic logistic growth. We consider a model wherein the possible reactions are birth events $N \rightarrow N + 1$, with a rate βN , and death events $N \rightarrow N - 1$, with rate δN^2 , where N is the population. We set $\beta = 2 \text{ s}^{-1}$ and $\delta = 0.001 \text{ s}^{-1}$. (a) An equation-free trajectory, which is derived through a combination of stochastic simulation runs (black points) and extrapolation (red line), and the deterministic solution (green broken line). The periods over which the exact simulations were performed are indicated by gray boxes. (b) A magnification of the sample path illustrates the procedure. (c) A schematic diagram of the equation-free method and its four stages. 1. Simulation. Exact stochastic simulation is executed. 2. Restriction. Coarse state variables are calculated based on the system state at the end of exact simulation. 3. Projection. The coarse variables are projected forward in time. 4. Lifting. A detailed system configuration is generated from the coarse state variables for the next exact simulation.

Figure 2.

Illustration of a model for the assembly of a linear aggregate. (a) The contact map of the model, which consists of six molecules A, B, C, D, and E considered in the model and their interactions. Each molecule has specific binding site(s), which are responsible for interaction with sites in other molecules. There are 21 possible complexes in this model. (b) Shorthand notation to denote the configurations of a single molecule. (c) First-degree moieties using the short-hand notation. The shaded boxes identify redundant information because populations of these moieties can be deduced from the populations of the unshaded moieties. (d) Second-degree moieties.

Figure 3. Moiety variables for a model for early events in EGFR signaling. (a) The contact map of the model. (b) The first-degree moieties. The moiety variables can be defined as model outputs (i.e., observables) using the BioNetGen language (BNGL).⁷ Redundant information involving *EGF* and *SOS* molecules is removed, similar to the linear assembly model. (c) The second-degree moieties.

Figure 4.

Stochastic simulation of EGFR signaling according to the model of Blinov et al.⁸ with equation-free acceleration. For this model, the reaction network and the full list of reacting species can be generated from the model's rules using BioNetGen.⁹ To perform stochastic simulations, we implemented Gillespie's direct method¹⁰ with equation-free acceleration in problem-specific code. Each burst of stochastic simulation consisted of 4096 reaction events. At each event time, we calculated and recorded the values of the coarse-grained moiety variables. To estimate derivatives for these variables, we performed linear fits to their stochastic time courses. The resulting slopes were taken to be the time derivatives of the variables, which were used in projection via Euler's method. The time increment used in projection was chosen to be three times the duration of the prior burst of stochastic simulation.

Figure 5.

The efficiency of equation-free acceleration of stochastic simulation for the EGFR model with $K = 1, 2, 3$, and 4 relative to exact stochastic simulation (Gillespie's direct method). The simulations were independently executed in parallel (32 simultaneous threads) on a machine with 32 Intel Xeon CPUs (E5-2698 v3, 2.30 GHz).