Inverse Gillespie for inferring stochastic reaction mechanisms from intermittent samples

Ishanu Chattopadhyay[a,1], Anna Kuchina[b], Gürol M. Süel[b], and Hod Lipson[a,c]

[a]School of Mechanical and Aerospace Engineering and [c]Computing and Information Science, Cornell University, Ithaca, NY 14853; and [b]Division of Biological Sciences and Section of Molecular Biology, University of California, San Diego, La Jolla, CA 92093

Edited by Eric D. Siggia, The Rockefeller University, New York, NY, and approved June 19, 2013 (received for review August 21, 2012)

Gillespie stochastic simulation is used extensively to investigate stochastic phenomena in many fields, ranging from chemistry to biology to ecology. The inverse problem, however, has remained largely unsolved: How to reconstruct the underlying reactions de novo from sparse observations. A key challenge is that often only aggregate concentrations, proportional to the population numbers, are observable intermittently. We discovered that under specific assumptions, the set of relative population updates in phase space forms a convex polytope whose vertices are indicative of the dominant underlying reactions. We demonstrate the validity of this simple principle by reconstructing stochastic models (reaction structure plus propensities) from a variety of simulated and experimental systems, where hundreds and even thousands of reactions may be occurring in between observations. In some cases, the inferred models provide mechanistic insight. This principle can lead to the understanding of a broad range of phenomena, from molecular biology to population ecology.

automated inference | population dynamics | machine science | Hough transform

Data-driven reverse engineering of dynamical systems has been a problem of longstanding interest (1, 2). Classical modeling methods have often used deterministic approximations based on coupled ordinary differential equations, yet many natural processes involving populations of interacting agents have an important stochastic component. To be valid, deterministic approximations require large agent populations that effectively average out the macroscopic effect of stochastic fluctuations. Recent investigations into cellular dynamics have revealed situations where this approximation fails. Stochastic effects have been implicated in diverse phenomena ranging from cell differentiation to pathogen virulence (3), maintenance of circadian rhythms (4, 5), bistability (6), excitability (7, 8), regulation of intracellular protein levels (9), and also in predator–prey in ecosystems (10). Experimental evidence confirms that gene expression occurs in abrupt stochastic bursts (11), and intercellular stochastic variation impacts high-level outcomes such as development and aging (12). Researchers are now able to estimate the number of specific macromolecules within a cell (13, 14), prompting a need for tools that explicitly model stochastic phenomena.

Standard Forward Algorithm and the Inverse Problem

Gillespie's stochastic simulation algorithm (SSA) (10) is an accurate procedure for simulating time evolution of finite interacting populations in continuous time; defining a computational approach to simulate the intractable chemical master equation. SSA along with its generalizations (15, 16) underlies most state-of-the-art stochastic kinetic modeling techniques (17).

Although providing an elegant technique to simulate a completely specified Markov jump process (MJP) (18), as is required to simulate chemical interaction of discrete agents, it is not particularly clear how observed time series data may be incorporated in a principled manner in Gillespie's SSA. Thus, the inverse problem of inferring the structure and parameters of biochemical networks from observed expression data has remained largely unsolved.

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 establishes that the probabilities of individual reactions are possibly nonlinear functions of instantaneous population numbers; because 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 the identification process notoriously difficult. While the calibration problem, i.e., estimation of parameters, namely the reaction propensities, given a model structure and observed expression data has seen significant progress (19–22), there is little reported work on de novo structure identification.

The present work delineates a principle to reverse-engineer observed population time series for de novo structural identification along with estimation of reaction propensities (Fig. 1). We only need intermittent measurements of the system state represented as population counts of each participating species, and may skip many (on the order of thousands in some of our examples) reactions between successive measurements. We do have limitations, e.g., following Gillespie we assume our system is well-mixed, that reactions have constant propensities, and additionally, our approach requires the system to admit a probabilistic equilibrium, and even then not all reactions are guaranteed to be identifiable (see *Assumptions, Limitations, and Error Sources* for details). Even with these limitations, we successfully demonstrate de novo inference on both simulated data from synthetic systems, and on experimental data from ecological and biochemical scenarios. Also, from a computational viewpoint, our technique is significantly simpler than Markov chain Monte Carlo (MCMC) simulation-based parameter-estimation schemes, and is validated to produce parameter estimates close to reported MCMC calibration techniques (20, 23) (*SI Appendix*, section SI-9).

Key Insight

For a multispecies system evolving stochastically via a small number of underlying reactions occurring with approximately constant propensities, with the population counts of the interacting species observed intermittently, the set of relative update vectors has an identifiable geometric structure. The specific geometry of the set of relative updates is indicative of the dominant reactions driving the system, and can be robustly identified under missed observations, measurement noise, and even limited stochastic fluctuations of the reaction propensities. Using Gillespie's original formulation of stochastic reaction systems, and some additional assumptions, we claim that the set of relative population update vectors defines a convex bounded polytope. More importantly, each vertex of this polytope has a direction cosine which is the same as that of the population update realized by a driving reaction. Furthermore, this alignment is independent, up to 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

Author contributions: I.C., A.K., G.M.S., and H.L. designed research; I.C., A.K., G.M.S., and H.L. performed research; I.C., A.K., G.M.S., and H.L. contributed new reagents/analytic tools; I.C., A.K., G.M.S., and H.L. analyzed data; and I.C. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

[1]To whom correspondence should be addressed. E-mail: ishanu.chattopadhyay@cornell.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1214559110/-/DCSupplemental.

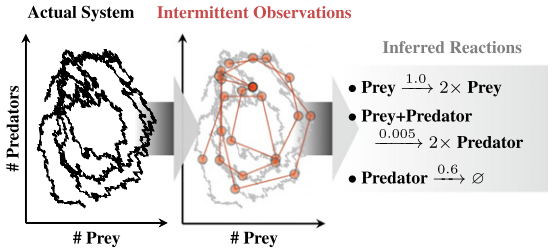

Fig. 1. Reverse-engineering a reaction model from observable "way-points" on the dynamical trajectory. (*Left*) Trajectory over time of two populations (e.g., prey and predator). Whereas the populations interact frequently, prey and predator quantities are only observed intermittently (red circle way-points). From these dots alone (*Center*), we reconstruct the most likely generative process (*Right*) underlying the data, leading to mechanistic insight and predictions. Actual data and automatically inferred system shown.

reactions, may be identified from the observed population time series alone, with no a priori system knowledge; thus defining a unique approach for de novo structure identification.

A reaction system G, as originally formalized by Gillespie, is a triple $(\mathcal{R},\mathcal{L},c)$. $\mathcal{R}$ and $\mathcal{L}$ are $R\times P$ integer-valued matrices (R: number of reactions, P: number of species) defining the reactions:

$$\begin{aligned} r_{11}X_1+\cdots+r_{1j}X_j+\cdots &\xrightarrow{c_1} \ell_{11}X_1+\cdots+\ell_{1j}X_j+\cdots \\ \vdots &\qquad\vdots \\ r_{i1}X_1+\cdots+r_{ij}X_j+\cdots &\xrightarrow{c_i} \ell_{i1}X_1+\cdots+\ell_{ij}X_j+\cdots \\ \vdots &\qquad\vdots \end{aligned}$$

where r_{ij},ℓ_{ij} are elements of $\mathcal{L},\mathcal{R}$ respectively, and $c_i, i=1,\ldots,R$ are the reaction propensities. G may be simulated in the forward direction using Gillespie's SSA once the initial population counts of each species is known. Given a value of the observation gap Δ, we define the density function $\psi^{\Delta}:(\mathbb{N}\cup\{0\})^P\times(\mathbb{N}\cup\{0\})^P\to[0,1]$, such that from the current population vector or state q, the probability that the next observed state is q', given $q'\neq q$, is $\psi^{\Delta}(q,q')$. In the limit $\Delta\to 0^+$, every reaction is observable; $\psi^0(q,q')$ is the probability of realizing q' from q by a single reaction. For $\Delta>0$, $\psi^{\Delta}(q,q')$ is the probability of observing q' next as a result of an unspecified number of unobserved reactions occurring within Δ. Finally, given $\eta>0$, we define the reaction polytope (RP):

$$\mathrm{RP}=\mathrm{Conv}^{+}\left(\bigcup_{q\in X}\{(q'-q):\psi^{\Delta}(q,q')\geqq\eta\}\right), \qquad [1]$$

where $q'\in(\mathbb{N}\cup\{0\})^P$ is the state observed immediately after q, $\mathrm{Conv}^{+}(\cdot)$ denotes the convex hull + interior, and $X\subset(\mathbb{N}\cup\{0\})^P$ is a finite set of population vectors visited by the system.

For a reaction system $G=(\mathcal{L},\mathcal{R},c)$, the species-specific population updates brought about by the driving reactions define the reaction directions, which are in fact the direction cosines of the rows of the matrix $\mathcal{F}=\mathcal{R}-\mathcal{L}$. Under our assumptions, the RP is necessarily a convex bounded polytope (*SI Appendix*, sections SI-1.4 and SI-2). Additionally, for any choice of η, X and up to a limiting magnitude of the observation gap Δ, each vertex of the RP necessarily coincides with a reaction direction (*SI Appendix*, Proposition SI-2). As a consequence, we may estimate the direction cosines of the RP vertices and hence the directions of the dominant reactions by identifying the bounding hyperplanes for the set of observed relative updates. Of course, the presence of noise and the finiteness of the observed time series data implies that simple calculation of the bounding hyperplanes, e.g., by taking the convex hull, would lead to spurious RP vertices. We solve this problem via the application of a Hough transform (HT) (24) on the set of relative updates (*SI Appendix*, section SI-2), and by carrying out robust identification of the bounding hyperplanes in the Hough parameter space. HT maps each observed point in the space of relative update vectors to a sinusoidal hypersurface (a sinusoidal curve in 2D) in the Hough space, and the intersection of these hypersurfaces (lines in 2D) indicate hyperplanes in the original space. The robustness of the approach (*SI Appendix*, section SI-2.3) arises from the established statistical properties of the HT estimator that is shown to be strongly consistent, with a finite-sample-replacement breakdown point (25, 26) close to 50%. Whereas traditionally the HT has been used to detect straight edges in 2D images in computer vision

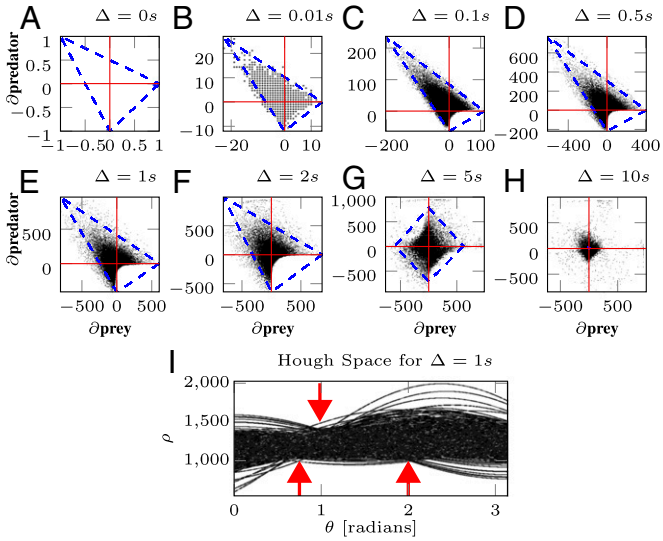

Fig. 2. Reaction polytopes. (*A–H*): RP geometry for the standard LV system with propensities (1, 0.005, 0.06) and with different observation gaps. As Δ is increased beyond 2 s, the correct polytopal geometry is lost. We argue in *SI Appendix*, section SI-1.4 (see also *SI Appendix*, Remark SI-1.7) that this is due to the significant variation of the reaction probabilities within the observation gap if the latter is large. (*I*) Illustration of the Hough parameter space corresponding to $\Delta=1$ s. Each relative update vector (e.g., a point in *E*) is mapped to a sinusoid in the Hough space, which represents the parameters of all straight lines passing through that point. We argue that the kinks (gradient discontinuities) in the curves that bound the nonzero Hough accumulator bins (pointed out by the red arrows) correspond to the edges of the RP, which in turn correspond to the direction cosines of the population updates brought about by the dominant driving reactions.

applications, we generalize it to P dimensions, and the intersection of the Hough hyperplanes (Hough lines for $P=2$) indicates the identifiable reaction directions (Fig. 2) with a high probability (approaching 1 exponentially fast with the length of the observed time series; *SI Appendix*, Proposition SI-2.2).

To see an example of the identified RP, bounded by Hough lines, examine the Lotka–Volterra (LV) predator–prey dynamical system, shown in Fig. 1. The LV system is defined by three simple reactions, namely: prey multiplication, predation by which predator multiplies at the expense of the prey, and predator death. We constructed the RP for simulated LV data with a range of different observation gaps. The result reveals a triangular polytope (Fig. 2 *A–H*). The polytope edges in Fig. 2*E* ($\Delta = 1s$) intersect at the coordinates $(250,0), (0,200), (-970,975)$, implying three reactions with directions 0°, 270°, and ∼135° which correspond to the update vectors $(1,0), (0,-1)$, and $(-1,1)$, respectively.

Not all reaction directions are guaranteed to show up as RP vertices (*SI Appendix*, section SI-2.5). The ones that do are the dominant reactions. We must still infer the exact stoichiometries of the reactants and products for each inferred direction. We determine species-specific upper bounds on the entries of $\mathcal{L}$ (*SI Appendix*, section SI-3), which then yield a finite set of possibilities for the stoichiometries. This results in a set of $(\mathcal{L}, \mathcal{R})$ pairs as plausible reaction structures for the system. We consider all such structures, find reaction-specific propensities in each case, and evaluate the complete models for error vs. complexity tradeoff on a Pareto front (*SI Appendix*, section SI-4.2).

To estimate propensities, we assume that there exists a population vector E with zero expected update, i.e., a probabilistic equilibrium. This implies the existence of a nonnegative nontrivial vector $\wp$ satisfying (*SI Appendix*, section SI-5.1)

$$\wp(\mathcal{R} - \mathcal{L}) = 0. \qquad [2]$$

Stochastic evolution, in general, traces out different trajectories for the same initial conditions. However, in an ε-neighborhood of a probabilistic equilibrium $(\mathcal{N}_\epsilon(E))$, a nearly zero expected update implies that the occurrence of reactions can be modeled as independent trials in a multinomial distribution, with each outcome independently drawn from the set of the R inferred reactions. This independence condition, combined with the observed variance of the updates in $\mathcal{N}_\epsilon(E)$, allows us to solve for the expected propensities. We show (*SI Appendix*, section SI-5) that there exists a consistent estimator $\widetilde{C}_r$ for the propensity of the rth reaction, whose expected value $\widetilde{c}_r$ (i.e., the propensity estimate) may be computed as

$$\widetilde{c}_r \Delta = \frac{\frac{1}{M}\sum_{j=1}^{M}\sum_{i=1}^{P} x_{ij}^2}{\left(\sum_{k=1}^{R}\left(\sum_{j=1}^{P}\left(\mathcal{R}_{kj} - \mathcal{L}_{kj}\right)^2\right)\wp_k\right)} \times \frac{\wp_r}{\prod_{j=1}^{P}\binom{E_j}{\mathcal{L}_{rj}}}, \qquad [3]$$

where x_{ij} is the population update for the ith species for the jth update observed to have occurred from within $\mathcal{N}_\epsilon(E)$, M is the total number of observed updates, and $\wp$ is an equilibrium probability vector (*SI Appendix*, section SI-5). Additionally, the coefficient of variation $\mathcal{V}_c$ for the estimator (i.e., the ratio of SD to mean), which quantifies the uncertainty in the estimate, is given by

$$\mathcal{V}_c = \sqrt{\frac{\sum_{j=1}^{M}\sum_{i=1}^{P} x_{ij}^4}{\left(\sum_{j=1}^{M}\sum_{i=1}^{P} x_{ij}^2\right)^2} - \frac{1}{M}}, \qquad [4]$$

which is identical for each reaction, and is independent of the equilibrium location and the equilibrium reaction probabilities $\wp$. We also work out explicit corrections arising from measurement noise (*SI Appendix*, section SI-6). Note that if the left null space of $(\mathcal{R} - \mathcal{L})$ has dimension greater than 1, then $\wp$ could be nonunique. In such cases, we sample the set of feasible $\wp$ vectors, and as before, tradeoff error-vs.-complexity of the resulting models on a Pareto front.

The observation gap Δ is the inverse frequency at which experimental observations are made (*SI Appendix*, Remark SI-5.1), and is expected to be known. Without knowing Δ we can only infer $c_r\Delta$, and can only meaningfully compare the observed and simulated phase spaces near probabilistic equilibria. Because a probabilistic equilibrium has zero expected update for any Δ, the distortion of the phase space in a small neighborhood of any such equilibrium is small. Thus, the deviation of the phase space in an equilibrium neighborhood reflects the modeling error independent of the observation gap. In contrast, comparing the phase spaces away from equilibria would require us to simulate the inferred model with Δ set to a value close to that used in the original experiment.

Computational Steps

Our inference algorithm consists of six sequential steps (see *SI Appendix*, section SI-8 for complete pseudocode). In step 1, we

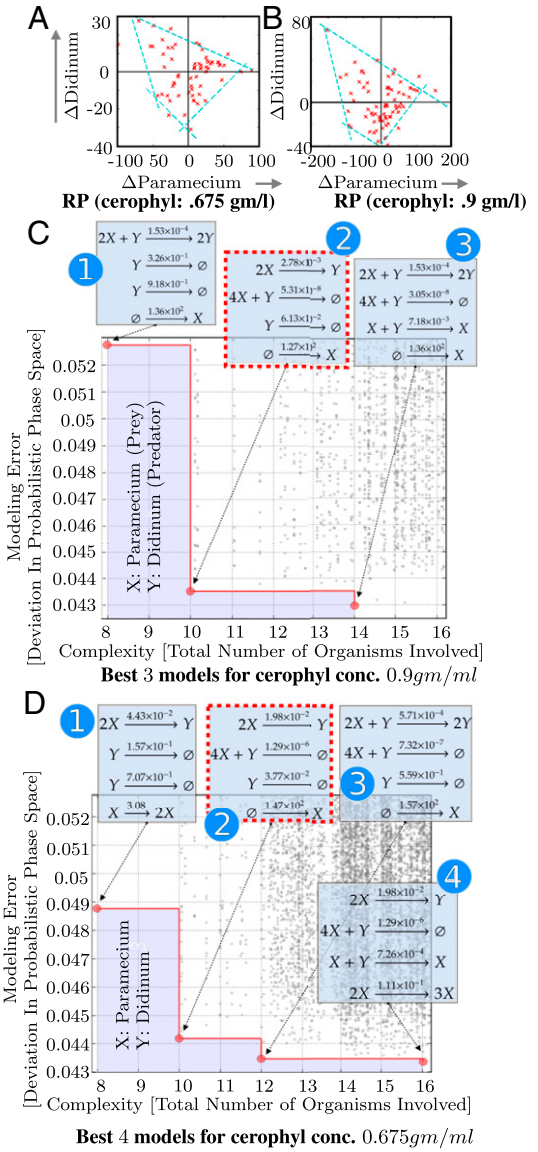

Fig. 3. Inferring *Paramecium–Didinum* cohabitation dynamics from experimental data (30). (*A* and *B*) RP for different cerophyl concentrations; (*C* and *D*): Pareto fronts. Model 2 is both simple and accurate, and occurs in both cases with different propensities.

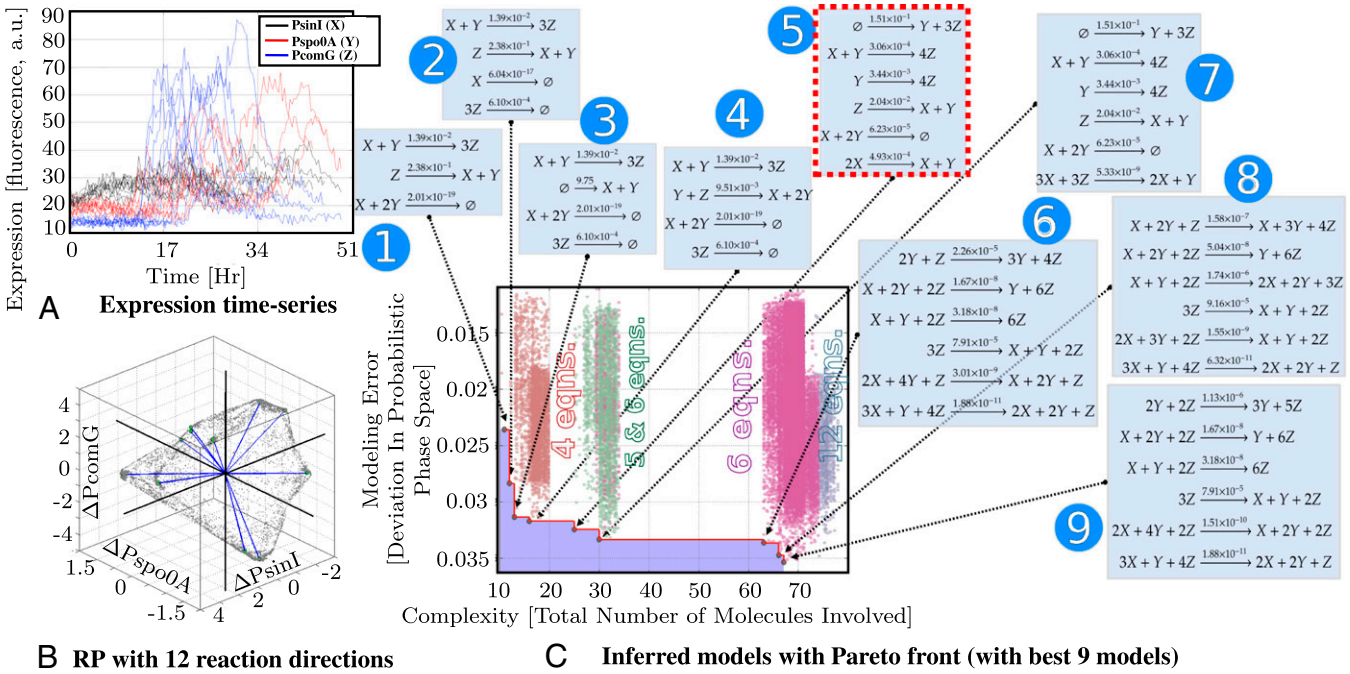

Fig. 4. Reverse-engineering competence response of B. subtilis. (*A*) Expression time series (for 10 cells) collected for promoters for genes sinI, spo0A, and comG (denoted as PsinI, Pspo0A, PcomG, respectively) implicated in competence. (*B*) Computed RP with 12 vertices, shown with blue line from the origin. Red dots are potential local depressions on the surface, which were rejected to be global vertices in the Hough analysis. (*C*) Pareto front with models generated with different number of reactions, integer coefficients, and reaction probability vectors at equilibrium. Model 5 offered surprising mechanistic insights, while vindicating known facts.

compute the optimal neighborhood radius (ϵ^*) (*SI Appendix*, section SI-5.5) as the one leading to the smallest coefficient of variation $\mathcal{V}_c$. (We need this radius to quantify what constitutes a neighborhood of any given state vector.) The location of a probabilistic equilibrium is identified as the population vector from which we have a near-zero expected update. Also, the variance of the observed population updates in the neighborhood of the identified equilibrium is computed. The expected update $U_{\epsilon^*}(v)$ from every point v in the discretized space of population vectors is computed as well by taking the average update from within a neighborhood of v. Step 2 estimates a species-specific upper bound for $\mathcal{L}$ entries. Step 3 applies HT to the set of relative updates, and outputs a set of $(\mathcal{R}, \mathcal{L})$ pairs representing possible model structures. For each $(\mathcal{R} - \mathcal{L})$ so obtained, we next estimate (step 5) the reaction probabilities producing zero expected update (*SI Appendix*, section SI-5.1). We implement this step by computing a basis $\mathcal{B}$ for the left null space for $(\mathcal{R} - \mathcal{L})$, and then searching for vectors in the span of $\mathcal{B}$ with nonnegative entries. No such $\wp$ may exist, which would render the problem unsolvable within our framework. In step 6 we evaluate the modeling error, and the descriptional complexity of the inferred models. We use the sum of the entries in the $\mathcal{L}$ and $\mathcal{R}$ as a measure of complexity (alternate measures of complexity may be used as well). Next we simulate each inferred model G (assuming $\Delta = 1$ if it is unknown) using Gillespie's SSA, and compute the optimal neighborhood radius ϵ', the expected updates $U'_{\epsilon'}$, and equilibrium E' as described in step 1. (Note: in step 1, we found these quantities, e.g., the optimal neighborhood ϵ_s, for the observed data; here we are finding them for simulated data from one of the inferred models.) Modeling error $(\mathcal{E}(G))$ in the vicinity of equilibria is then estimated using

$$\mathcal{E}(G) = \sum_{v \in \mathcal{N}_{\epsilon^*}(E)} \left\| U'_{\epsilon'}(v) - U_{\epsilon^*}(v) \right\| + \sum_{v \in \mathcal{N}_{\epsilon'}(E')} \left\| U'_{\epsilon'}(v) - U_{\epsilon^*}(v) \right\| \quad [5]$$

The model-specific error and complexity estimates can then be used for tradeoffs on a Pareto front.

Use of the Pareto Front. As in any machine learning task, there is a natural tradeoff between error and complexity, and we eliminate models that are not maximally accurate for a particular complexity value. This leads to only a few models on the error-complexity Pareto front, as shown in Figs. 3 and 4. An expert can then consider these models and draw insight based on additional domain knowledge.

The complexity measure dictates which models get reported on the Pareto front; thus, specific applications may necessitate more discriminative estimates.

Assumptions, Limitations, and Error Sources

Our approach makes the following assumptions:

i) Well-mixed reactants, i.e., absence of spatial inhomogeneity.

ii) Reaction propensities are constant, i.e., independent of reaction time, and instantaneous population vectors.

Even if these conditions are satisfied approximately, our inferred models are quite accurate (*SI Appendix*, sections SI-5.4 and SI-6.3).

iii) The system admits a probabilistic equilibrium (*SI Appendix*, Lemma SI-1.1). The notion of a probabilistic equilibrium is weaker than that of an equilibrium in a dynamical system. A system need not be static at a probabilistic equilibrium; only the expected change needs to be zero. This is realized with reaction probabilities that cause updates from individual reactions to cancel out. Thus, we cannot identify models for systems that do not exhibit any recurrent behavior, e.g., a system with a single reaction $X \to \varnothing$ (we can identify the reaction direction, but not the propensity).

iv) There is an upper bound on the observation gap beyond which the polytopal geometry of the RP is incorrect or lost (*SI Appendix*, Remark SI-1.7). We assume that Δ is smaller than this bound. A minor assumption is that Δ is approximately constant and known. Without Δ we can only estimate $c_r\Delta$ (Eq. **3**).

It is not guaranteed that every reaction will be identified, e.g., reactions with the same direction cosine cannot be distinguished. However, the algorithm approximates the observed dynamics with the reactions it can infer (*SI Appendix*, section SI-12).

In molecular-biology observables are concentration of marker proteins, and it is difficult to map these values back to exact population numbers (*SI Appendix*, section SI-7).

As Δ is increased, we expect a progressive degradation of the RP geometry (Fig. 2 *A–H*); e.g., in Fig. 2*H* with $\Delta = 5s$, the RP vertices appear incorrectly on the coordinate axes, illustrating the existence of an upper bound on the acceptable magnitude of Δ.

For a small number of species, HT-based RP identification scales easily with increasing complexity of the systems, both in terms of increasing number of reactions and increasing number of entities involved in the reactions (i.e., sum of the integer coefficients in the model), as shown in *SI Appendix*, Fig. S-7. Whereas the theory itself has been developed for an arbitrary number of species, classical HT implementations have a high memory overhead in higher dimensions (although the number of parameters increases linearly, the number of cells in the discretized parameter space increases exponentially). Notably, some recent breakthroughs have reported HT schemes that have polynomial complexity in the number of dimensions (*SI Appendix*, section SI-2.4).

Because we are looking for integer coefficients, small errors in the reaction directions have no effect on the solution. Measurement errors do however affect the propensity estimation (*SI Appendix*, section SI-6), and we calculate explicit corrections that may be applied if the variance of the noise is known.

In the vast majority of problems in molecular biology, we observe the expression of only a few key regulators, whereas such systems typically consist of hundreds of distinct species with complex interaction networks. For example, in our experimental investigation into the competence dynamics of *Bacillus subtilis* (see below), we simultaneously measured the expression of *sinI*, *spo0A*, and *comG* promoters. However, it is well known that PsinI, Pspo0A, and PcomG are only key regulators in a complex network of hundreds of genes, and do not interact in the same sense as reacting molecules combine to produce products. Nevertheless, the "reactions" we infer from the expression data of PsinI, Pspo0A, and PcomG do make physical sense (see the discussion in the next section) and represent high-level abstractions of key processes. We give examples of several simulated systems with unobserved interacting species (*SI Appendix*, sections SI-10 and SI-11), and show that the inferred models can correctly identify the key roles played by the observable species.

Application Examples

Synthetic Datasets. We validated our approach on a range of simulated and experimental systems. Simulated data included several variations of the LV system (systems 1 and 2 in Table 1), synthetic data from simulated transcription–translation dynamics (system 3 in Table 1; also Fig. 2*I*), and on the Oregonator (27) as a simple model of the Belousov–Zhabotinsky oscillator (system 4 in Table 1). The reaction structure was recovered correctly in each case (there is an obvious "correct" choice from the Pareto front for these specific systems, *SI Appendix*, section SI-13), and the inferred propensities were within 4% of the true values. The observation gap used varied between the systems, and the maximum value at which inference was possible is tabulated in Table 1.

Challenge Problems. Two datasets, comprising 425 and 315 observations, respectively, from the LV system are included, as benchmark challenges for competing approaches (*SI Appendix*, section SI-15).

Experimental Datasets. We first analyzed paramecium–didinum prey–predator cohabitation (28, 29) (paramecium denoted as species X, and didinum denoted as species Y). This is a classic protozoan experiment with *Paramecium aurelia* as prey and *Didinum nasutum* as predator (30), yielding time series of twice-daily counts of predator and prey abundance over several sustained population cycles

Table 1. Applications: Synthetic and experimental systems

System Description and inferred model	Performance
1. Lotka-Volterra (Standard) $X \xrightarrow{1.0} 2X$ $X+Y \xrightarrow{0.005} 2Y$ $Y \xrightarrow{0.6} \varnothing$ Inferred $X \xrightarrow{1.0042} 2X$ $X+Y \xrightarrow{0.005} 2Y$ $Y \xrightarrow{0.6025} \varnothing$	$\Delta = 2s$ $\mathcal{V}_c \simeq 2\%$ $e^{\dagger} \simeq 0.4\%$
2. Lotka-Volterra (Modified) $X \xrightarrow{0.5} 2X$ $2X+2Y \xrightarrow{.0001} 4Y$ $Y \xrightarrow{0.3} \varnothing$ Inferred $X \xrightarrow{0.494} 2X$ $2X+2Y \xrightarrow{.00009} 4Y$ $Y \xrightarrow{0.311} \varnothing$	$\Delta = 2s$ $\mathcal{V}_c \simeq 2\%$ $e^{\dagger} \simeq 3.7\%$
3. Transcrp. transl. (DNA $\leftrightarrow$ mRNA $\longrightarrow$ protein)[‡] $D \xrightarrow{1.0} D+M$ $M \xrightarrow{0.5} M+P$ $M \xrightarrow{1.0} \varnothing$ $P \xrightarrow{0.5} \varnothing$ Inferred $\varnothing \xrightarrow{122.4} M$ $M \xrightarrow{0.502} M+P$ $M \xrightarrow{1.03} \varnothing$ $P \xrightarrow{0.503} \varnothing$	$\Delta = 1s$ $\mathcal{V}_c \simeq 2\%$ $e^{\dagger} \simeq 2.3\%$ ($D = 120$ constant conc.)
4. Belousov-Zhabotinsky Oscillator (Oregonator)[¶] $X_1+Y_2 \xrightarrow{.656} Y_1$ $Y_1+Y2 \xrightarrow{5.22} Z_1$ $X_2+Y_1 \xrightarrow{3.27} 2Y_1+Y_3$ $2Y_1 \xrightarrow{1.05} Z_2$ $X_3+Y_3 \xrightarrow{8.2} Y_2$ Inferred $Y_2 \xrightarrow{990} Y_1$ $Y_1+Y2 \xrightarrow{5.02} \varnothing$ $Y_1 \xrightarrow[\times 10^3]{49} 2Y_1+Y_3$ $2Y_1 \xrightarrow{0.98} \varnothing$ $Y_3 \xrightarrow{12290} Y_2$	$\Delta = 5\mu s$ $\mathcal{V}_c \simeq 2\%$ $e^{\dagger} \simeq 6.7\%$ ($X_i = 15\times 10^3$)
5. Experimental Hudson Bay Data (X:Hare, Y:Lynx) $X \xrightarrow{0.464} 3X$ $X+Y \xrightarrow{0.026} 3Y$ $Y \xrightarrow{1.816} \varnothing$	$\mathcal{V}_c \simeq 6.1\%$ $\mathcal{E} \simeq 2.7\%$
6. Experimental Paramecium Didinum co-habitation (X:Paramecium, Y:Didinum) a. Cerophyl : 0.675 *gm/l* $2X \xrightarrow{0.000278} Y$ $4X+Y \xrightarrow[\times 10^{-8}]{5.31} \varnothing$ $Y \xrightarrow{0.0613} \varnothing$ $\varnothing \xrightarrow{127} X$ b. Cerophyl : 0:9 *gm/l* $2X \xrightarrow{0.000198} Y$ $4X+Y \xrightarrow[\times 10^{-6}]{1.29} \varnothing$ $Y \xrightarrow{0.0377} \varnothing$ $\varnothing \xrightarrow{147} X$	a. $\mathcal{V}_c \simeq 4.7\%$ $\mathcal{E} \simeq 5\%$ b. $\mathcal{V}_c \simeq 1.6\%$ $\mathcal{E} \simeq 5\%$
7. Experimental Cellular Differentiation Dynamics of *B. subtilis* (X: PsinI, Y: Pspo0A, Z: PcomG) $\varnothing \xrightarrow{0.151} Y+3Z$ $X+Y \xrightarrow{0.00031} 4Z$ $Y \xrightarrow{0.0034} 4Z$ $Z \xrightarrow{0.02} X+Y$ $X+2Y \xrightarrow{0.000062} \varnothing$ $2X \xrightarrow{0.00049} X+Y$	$\mathcal{V}_c \simeq 5.2\%$ $\mathcal{E} = 3.5\%$

[†]Percentage error in identified propensities (simulated systems).
[‡]Since D is constant at 120, the best we can do is identify $\varnothing \rightarrow M$ with propensity 1.0×120. The identified value is close (122.4)
[§]Similarly, the propensities of the first, third and fourth reactions are scaled up by $\simeq 15000$, and $\forall i X_i, Z_i$ is missing in the model

(experimental runs spanned 25 – 30 days each). We analyzed the two longest time series observed, measured under with cerophyl concentrations of 0.675 and 0.9gm/l, respectively.

Inferred Pareto fronts are shown in Fig. 3 *C* and *D*, respectively. The inferred models are similar to the standard LV systems as expected. More importantly, model 2 from Fig. 3*C* and model 2 from Fig. 3*D* have identical reactions, with different propensities. The fourth reaction $\varnothing \to X$ proceeds significantly faster for higher cerophyl concentration (1.27×10^2 in Fig. 3*C*) against 1.47×10^2 in Fig. 3*D*; higher cerophyl concentration gives more nutrition to the protozoa allowing them to multiply faster.

Abundance of food also correlates with the lower rates of population decay for both species (reactions 2 and 3 in the respective models). The large coefficient on paramecium (X) in the decay reaction $4X + Y \to \varnothing$ implies that at low protozoa concentrations the probability of this reaction dominates, bringing about rapid decay of protozoa population, resulting in fast extinction of both species (30).

The second experimental system we analyzed is cellular differentiation dynamics of the microbial model system *B. subtilis*, known to exhibit stochastic behavior at the single-cell level (31–34). Upon exposure to environmental stressors, individual *B. subtilis* cells can probabilistically differentiate into highly resilient dormant spores, or transiently enter a state of competence for uptake of extracellular DNA. Both processes require expression of Spo0A, which is the global transcriptional stress response regulator of *B. subtilis* (35). Spo0A governs transcription of several hundred genes, including *sinI*, whose expression releases inhibition of sporulation by SinR. Therefore, expressions of Spo0A and SinI are critical steps in the sporulation program (36). Furthermore, Spo0A expression removes inhibition on the transcriptional master regulator ComK, which activates expression of over 140 genes involved in competence (37). Among these is *comG*, which is crucial for extracellular DNA uptake. Measurement of *spo0A*, *sinI,* and *comG* promoters thus provides critical insight into stochastic differentiation dynamics of two competing cellular differentiation programs in single cells (38).

We applied our technique to simultaneously recorded expression data (Fig. 4*A*) of *spo0A*, *sinI,* and *comG* promoters (denoted Pspo0A, PsinI, and PcomG, respectively) in cells undergoing competence episodes followed by sporulation. The RP showed 12 distinct vertices, indicating 12 reactions (Fig. 4*B*). Some of these vertices occurred in clusters, and could be collapsed. We computed the reaction directions at different resolutions, and obtained models with number of inferred reactions ranging from 3 to 12, leading to the Pareto front shown in Fig. 4*C*.

Model 5, consisting of six reactions, revealed surprising insights into the complex cellular differentiation dynamics and relationship between sporulation and competence.

Reactions ($\varnothing \to \mathrm{Pspo0A} + 3\mathrm{PcomG}$, $2\mathrm{PsinI} \to \mathrm{PsinI} + \mathrm{Pspo0A}$, $\mathrm{PsinI} + 2\mathrm{Pspo0A} \to \varnothing$) capture the known requirement for entry into competence. Specifically, these reactions confirm that Spo0A expression, and thus at least a basal activity of the sporulation program, precedes competence initiation (35). Reactions ($\mathrm{PcomG} \to \mathrm{PsinI} + \mathrm{Pspo0A}$, $\mathrm{PsinI} + \mathrm{Pspo0A} \to 4\mathrm{PcomG}$) explain the controlled concentration of PcomG during competence. Also PcomG → PsinI + Pspo0A corroborates recent experimental evidence of competence episodes being often followed by sporulation (38). Finally, PsinI + 2Pspo0A → ∅ predicts that both Spo0A and SinI are repressed during competence. Thus, this inferred model recovers known dynamics, and poses nontrivial hypotheses for unknown interactions.

Conclusion

Under specific assumptions, certain observable geometric properties of the set of relative population updates in the phase space remain approximately conserved when an unknown number of reactions is skipped between aggregate population counts; this fact can be exploited for de novo inference of reaction structure and propensities with noisy and intermittent observations. Such de novo inference is crucial to analyzing large volumes of experimental data acquired via a gamut of emerging single-cell monitoring technologies, and may help elucidate fundamental principles by which biological elements integrate to exhibit complex behavior.

ACKNOWLEDGMENTS. This research was supported in part by the US Army Research Office Biomathematics program (W911NF-12-1-0499), the National Science Foundation (ECCS 0941561), and the Defense Threat Reduction Agency (HDTRA 1-09-1-0013).

1. King RD, et al. (2009) The automation of science. *Science* 324(5923):85–89.
2. Yule UG (1927) On a method of investigating periodicities in disturbed series, with special reference to Wolfer's sunspot numbers. *Philos Trans R Soc Lond* 226:267–298.
3. McAdams HH, Arkin A (1997) Stochastic mechanisms in gene expression. *Proc Natl Acad Sci USA* 94:814–819.
4. Forger DB, Peskin CS (2005) Stochastic simulation of the mammalian circadian clock. *Proc Natl Acad Sci USA* 102(2):321–324.
5. Mihalcescu I, Hsing W, Leibler S (2004) Resilient circadian oscillator revealed in individual cyanobacteria. *Nature* 430(6995):81–85.
6. Ferrell JE (2002) Self-perpetuating states in signal transduction: Positive feedback, double-negative feedback and bistability. *Curr Opin Cell Biol* 14(2):140–148.
7. Bratsun D, Volfson D, Tsimring LS, Hasty J (2005) Delay-induced stochastic oscillations in gene regulation. *Proc Natl Acad Sci USA* 102(41):14593–14598.
8. Edery I, Zwiebel LJ, Dembinska ME, Rosbash M (1994) Temporal phosphorylation of the Drosophila period protein. *Proc Natl Acad Sci USA* 91(6):2260–2264.
9. Sigal A, et al. (2006) Variability and memory of protein levels in human cells. *Nature* 444(7119):643–646.
10. Gillespie DT (1977) Exact stochastic simulation of coupled chemical reactions. *J Phys Chem* 81(25):2340–2361.
11. Zlokarnik G, et al. (1998) Quantitation of transcription and clonal selection of single living cells with beta-lactamase as reporter. *Science* 279(5347):84–88.
12. Finch CE, Kirkwood T (2000) Chance, development, and aging. *Biogerontology* 1(4):373–373.
13. Cai L, Friedman N, Xie XS (2006) Stochastic protein expression in individual cells at the single molecule level. *Nature* 440(7082):358–362.
14. Suel G (2011) Synthetic Biology, Part A: Methods for Part/Device Characterization and Chassis Engineering. *Methods in Enzymology*, ed Voigt, C (Academic, Waltham, MA) 1st Ed, Vol 497.
15. Gillespie DT (2001) Approximate accelerated stochastic simulation of chemically reacting systems. *J Chem Phys* 115:1716–1733.
16. Cao Y, Gillespie DT, Petzold LR (2007) Adaptive explicit-implicit tau-leaping method with automatic tau selection. *J Chem Phys* 126(22):224101.
17. Wilkinson D (2006) *Stochastic Models for System Biology*. Chapman and Hall/CRC Mathematical and Computational Biology Series (Taylor & Francis).
18. Opper M, Sanguinetti G (2007) *Variational Inference for Markov Jump Processes in NIPS 2007*, eds Platt JC, Koller D, Singer Y, Roweis ST (Curran Associates).
19. Boys RJ, Wilkinson DJ, Kirkwood TB (2008) Bayesian inference for a discretely observed stochastic kinetic model. *Stat Comput* 18(2):125–135.
20. Golightly A, Wilkinson DJ (2011) Bayesian parameter inference for stochastic biochemical network models using particle markov chain monte carlo. *Interface Focus* 1(6):807–820.
21. Milner P, Gillespie CS, Wilkinson DJ (2013) Moment closure based parameter inference of stochastic kinetic models. *Stat Comput* 23(2):287–295.
22. Amrein M, Künsch HR (2012) Rate estimation in partially observed Markov jump processes with measurement errors. *Stat Comput* 22(2):513–526.
23. Henderson DA, Boys RJ, Wilkinson DJ (2010) Bayesian calibration of a stochastic kinetic computer model using multiple data sources. *Biometrics* 66(1):249–256.
24. Hough PVC (1962) US Patent 3,069,654.
25. Huber P (2005) *Robust Statistics*. Wiley Series in Probability and Mathematical Statistics. Probability and Mathematical Statistics (Wiley, New York).
26. Dattner I (2009) Statistical properties of the Hough transform estimator in the presence of measurement errors. *J Multivariate Anal* 100(1):112–125.
27. Field R (1975) Limit cycle oscillations in the reversible Oregonator. *J Chem Phys* 63(2289):8.
28. Gause G (1935) Experimental demonstrations of Volterra's periodic oscillations in the numbers of animals. *J Exp Biol* 12:44–48.
29. Luckinbill, L (1972) *Coexistence in Laboratory Populations of Paramecium Aurelia and the Predator Didinium Nasutum* (Univ California, Los Angeles-Zoology).
30. Veilleux B (1979) An analysis of the predatory interaction between paramecium and didinium. *J Anim Ecol* 48:787–803.
31. Suel GM, Garcia-Ojalvo J, Liberman LM, Elowitz MB (2006) An excitable gene regulatory circuit induces transient cellular differentiation. *Nature* 440(7083):545–550.
32. Suel GM, Kulkarni RP, Dworkin J, Garcia-Ojalvo J, Elowitz MB (2007) Tunability and noise dependence in differentiation dynamics. *Science* 315(5819):1716–1719.
33. Maamar H, Raj A, Dubnau D (2007) Noise in gene expression determines cell fate in bacillus subtilis. *Science* 317(5837):526–529.
34. Cagatay T, Turcotte M, Elowitz M, Garcia-Ojalvo J, Suel G (2009) Architecture-dependent noise discriminates functionally analogous differentiation circuits. *Cell* 139(3):512–522.
35. Grossman AD (1995) Genetic networks controlling the initiation of sporulation and the development of genetic competence in Bacillus subtilis. *Annu Rev Genet* 29:477–508.
36. Stragier P, Losick R (1996) Molecular genetics of sporulation in bacillus subtilis. *Annu Rev Genet* 30:297–341.
37. Dubnau D (1999) Dna uptake in bacteria. *Annu Rev Microbiol* 53:217–244.
38. Kuchina A, et al. (2011) Temporal competition between differentiation programs determines cell fate choice. *Mol Syst Biol* 7:557.