Preprints of the 19th World Congress
The International Federation of Automatic Control
Cape Town, South Africa. August 24-29, 2014

Identification of Markov Chains From Distributional Measurements and Applications to Systems Biology ⋆

Anandh Swaminathan * **Richard M. Murray** *

* *California Institute of Technology, Pasadena, CA 91125 USA (e-mail: aswamina@caltech.edu).*

Abstract: In this paper, we are concerned with doing system identification for biological systems using time series distributional measurements and Markov chain models. By distributional measurements, we mean measurements provided by assays such as flow cytometry or FISH (fluorescence in situ hybridization) that allow us to make the same single cell measurements over a number of different cells. We focus here on the problem of estimating the transition probabilities of a Markov chain from distributional measurements. We set this problem up using Bayes' rule and make simplifying assumptions to reduce the problem to a non-convex optimization problem over finitely many variables. We propose methods for locally solving this non-convex optimization problem. For a special case, we discuss necessary and sufficient conditions for the Markov chain to be identifiable. Finally, we demonstrate our procedure on instructive toy examples as well as on simulated stochastic data for a genetic toggle switch. [Gardner et al. (2000)].

Keywords: system identification, systems biology, parameter estimation, convex optimization, bio control

1. INTRODUCTION

In this paper, we are concerned with doing system identification for biological systems using time series distributional measurements. Distributional measurements are measurements provided by assays such as flow cytometry or fluorescence in situ hybridization that allow us to take the same single cell measurements over a number of different cells at each time step. Here, we assume that the underlying model for each cell is a finite Markov chain (FMC). Therefore, our problem reduces down to identifying the transition probabilities of a FMC from distributional measurements.

Biological systems are intrinsically stochastic [Elowitz et al. (2002)], and stochastic effects can greatly influence the function of biological circuits [Eldar and Elowitz (2010)]. Thus, stochastic models are relevant in systems biology, and system identification for these stochastic models is relevant as well.

Stochastic chemical kinetics models [Gillespie (1977)] are a popular type of stochastic model in systems biology. The system identification problem for stochastic chemical kinetics models consists of estimating the reaction propensities for a set of known chemical reactions from observations. Many approaches including Lillacci and Khammash (2013) use Bayesian methods and sampling techniques to tackle this problem, while other approaches leverage subspace methods for system identification [Hori et al. (2013)]. Munsky and Khammash (2010) and Lillacci and Khammash (2011) are also good references on estimating reaction propensities from flow cytometry data. However, some of these methods require computationally expensive sampling while more computationally efficient methods tend to require perfect measurements of all relevant species. These methods also require a priori knowledge of the relevant chemical species and reactions.

Machine learning methods are the other popular approach to identifying stochastic models in biology. These models assume minimal prior knowledge about the system and fit a model to the data that provides scientific insight about the system. For example, in Neuert et al. (2013), the authors fit a discrete-state continuous-time Markov model to the yeast transcriptional response to osmotic stress. The results showed that gene expression could be switched on and off and came on in multiple steps and provided insight into the underlying dynamics. However, the states in this model are black box states and do not correspond to biological properties. Other machine learning approaches such as Sachs et al. (2005) and Husmeier (2003) use Bayesian network methods to reveal the structure of signaling pathways. While these methods elucidate the structure of signaling pathways, the work in Neuert et al. (2013) is a good example of why understanding the dynamics matters as well.

The method we develop in this paper tries to bridge the gap between these two approaches to biological system identification. Given realistic sensors and minimal prior information about our system of interest, our goal is to find a Markov model for the dynamics of a biocircuit in a single cell, where the states in the model correspond to different levels of RNA and protein. Our approach is different

⋆ This work was supported by NSF GRFP.

from methods to estimate the transition probabilities such as the Baum-Welch algorithm [Russell et al. (1995)] because our time series measurements are distributional measurements rather than measurements for an individual cell. Also, methods for noisy linear system identification work poorly because they assume Gaussian noise, and this assumption limits performance when data is sparse.

The paper is organized as follows. In Section II, we motivate the FMC identification problem in the context of systems biology and then reduce it to an optimization problem. In Section III, we discuss methods to solve this optimization problem and recover an estimate of the FMC transition probabilities. In Section IV, we consider necessary and sufficient conditions under which the FMC transition probabilities are correctly identifiable from distributional measurements in the special case of full observability and infinite sample size. In Section V, we consider instructive toy examples as well as a biological example. We conclude in Section VI.

2. MOTIVATION AND OPTIMIZATION PROBLEM SETUP

2.1 Motivation of Problem

The nature of the single cell data collected in systems biology experiments suggests the use of a finite Markov chain model to model and identify the dynamics of systems in single cells. Biological measurement techniques such as FISH or flow cytometry provide noisy data at discrete time intervals. This data is typically binned, so there is a finite number of outputs. As a result, the natural model to consider for modeling this data is a finite Markov chain. Before delving into our specific problem formulation, we discuss some preliminaries on Markov chains.

2.2 Preliminaries

For a good review of Markov chains, see Levin et al. (2009). A finite Markov chain is a discrete time stochastic process on a finite state space Ω. Finite Markov chains must satisfy the Markov property, which states that given the current state, the next state is independent of the past. Specifically, if X_t is a random variable denoting the state of the chain at time step $t \in \mathbf{Z}_+$, then

$$\mathbf{P}(X_{t+1}|X_t) = \mathbf{P}(X_{t+1}|X_t, X_{t-1}, \ldots, X_0).$$

Therefore, letting $|\Omega| = n$, we can define a transition matrix $P \in \mathbf{R}_+^{n\times n}$ such that

$$\mathbf{P}(X_{t+1} = y|X_t = x) = P(y, x),$$

where $P(y, x)$ denotes the (y, x)-th entry of P. Additionally, we let the vector $\mathbf{x}_t \in \mathbf{R}_+^n$ denote the probability distribution of X_t where $\mathbf{P}(X_t = j) = \mathbf{x}_t(j)$. Then, we see that

$$P\mathbf{x}_t = \mathbf{x}_{t+1} \quad t \in \mathbf{Z}_+. \tag{1}$$

We refer to (1) as the dynamic constraint as it enforces that the probability distributions evolve according to the transition matrix.

Both the transition matrix P and state distributions $\mathbf{x}_t$ are confined to compact, convex sets. The transition matrix P must be column stochastic. Similarly, each $\mathbf{x}_t$ must be a stochastic vector. We can describe these constraints through inequalities by stating

$$\begin{aligned} P^*\mathbf{1} &= \mathbf{1} \\ P &\geq 0 \\ \mathbf{1}^*\mathbf{x}_t &= 1 \quad t \in \mathbf{Z}_+ \\ \mathbf{x}_t &\geq 0 \quad t \in \mathbf{Z}_+. \end{aligned} \tag{2}$$

The inequalities here are element wise and $\mathbf{1}$ refers to a column vector of ones of the appropriate dimension. We refer to equation (2) as the stochastic feasibility constraint. We are now ready to consider the problem of doing system identification of finite Markov chains from distributional measurements.

2.3 Problem Setup

Suppose that each cell in a large culture contains a biocircuit that is modeled by a finite Markov chain and that each cell is identical and evolves independently of all other cells. Then we have an infinite pool of Markov chains with transition matrix P and initial state distribution $\mathbf{x}_0$. The transition matrix P and initial distribution $\mathbf{x}_0$ are hidden, and we desire to estimate P. To do so, we run the chain for $T-1$ time steps and at each time step, we draw M chains from the pool and measure the state of each. We form a *data matrix* $D \in \mathbf{Z}_+^{n\times T}$, where the (i, j)-th entry $\mathbf{D}_{ij}$ is the number of sampled chains observed to be in state i at time step j.

We then want to calculate the maximum a posteriori (MAP) estimate of P having observed the data D. The principle of maximum entropy [Jaynes (1957)] tells us that the prior distribution over both P and the $\mathbf{x}_t$'s should be uniform. Applying Bayes' rule, we get

$$\begin{aligned} p(P|D) &\propto \int p(D|P, \mathbf{x})p(P)p(\mathbf{x})d\mathbf{x} \\ &\propto \int p(D|P, \mathbf{x})d\mathbf{x}, \end{aligned} \tag{3}$$

where $p(P|D)$ is the posterior probability density function (PDF) of P given the data D, $p(P)$ and $p(\mathbf{x})$ are uniform prior distributions, and $p(D|P, \mathbf{x})$ is the likelihood function.

To avoid computing the difficult integral in (3), we further approximate the posterior distribution as being peaked at one value of P and $\mathbf{x}$, which allows us to maximize the likelihood function over P and $\mathbf{x}$.

The final step in setting up the optimization problem is defining the likelihood function $p(D|P, \mathbf{x})$. First, we let $\mathbf{d}_0, \ldots, \mathbf{d}_{T-1}$ be the columns of D. Then, since we are drawing cells from an infinite population, we do not need to consider issues of replacement, so $\mathbf{d}_j$ is multinomially distributed with parameter $\mathbf{x}_j$. Then, using the multinomial probability mass function, the likelihood can be written down as

$$p(D|P, \mathbf{x}) = \prod_{j=0}^{T-1} \left(\frac{M!}{\mathbf{d}_j^1! \ldots \mathbf{d}_j^n!} \prod_{i=1}^{n} (\mathbf{x}_j^i)^{\mathbf{d}_j^i} \right), \tag{4}$$

where $\mathbf{d}_j^i$ refers to component i of $\mathbf{d}_j$, and $\mathbf{x}_j^i$ refers to component i of $\mathbf{x}_j$.

Note that this likelihood function is only valid for combinations of P and $\mathbf{x}$ that satisfy the stochastic constraint (2) and dynamic constraint (1). Taking the logarithm of the likelihood function (4) and dropping terms that only depend on the data, we arrive at the log-likelihood function $\ell(\mathbf{x})$, which is given by

$$\ell(\mathbf{x}) = \sum_{j=0}^{T-1} \sum_{i=1}^{n} \mathbf{d}_j^i \log \mathbf{x}_j^i. \tag{5}$$

We can then finally write down the following optimization problem, which is the full non convex optimization problem that we want to solve.

$$\begin{array}{lll} \underset{P,\mathbf{x}_0,\mathbf{x}_1,\ldots}{\text{minimize}} & -\ell(\mathbf{x}) & \\ \text{subject to} & P^*\mathbf{1} = \mathbf{1} & \\ & P \geq 0 & \\ & \mathbf{1}^*\mathbf{x}_t = 1 & t = 0,\ldots,T-1 \\ & \mathbf{x}_t \geq 0 & t = 0,\ldots,T-1 \\ & P\mathbf{x}_t = \mathbf{x}_{t+1} & t = 0,\ldots,T-2. \end{array} \tag{6}$$

By solving the optimization problem given by (6), we take the optimal value of P to be our estimate of the transition matrix. However, this problem is not convex because of the bilinear dynamic constraint.

Also, it is straightforward to add in affine constraints on P if we have information on the structure of P. We can also generalize the optimization problem to a case where we cannot observe the exact state of the chain and different states provide the same output. However, in the interest of clarity and space, these considerations are omitted in this paper.

3. SOLUTION METHODS TO OPTIMIZATION PROBLEM

In this section, we propose two methods for handling the non convex optimization problem given by (6). The first method simply offloads the problem to the fmincon function provided by MATLAB, while the second method utilizes the alternating direction method of multipliers (ADMM) [Boyd (2010)].

3.1 Direct Solution

In this approach, we note that there are only two variables that are effectively free in the model, the transition matrix P, and the initial state distribution $\mathbf{x}_0$. Once these two variables are specified, the distribution at each future time step is given by $\mathbf{x}_t = P^t\mathbf{x}_0$. Thus, we can rewrite the non-convex optimization problem (6) as an optimization problem over only P and $\mathbf{x}_0$. This yields the reduced non-convex optimization problem (7).

$$\begin{array}{lll} \underset{P,\mathbf{x}_0}{\text{minimize}} & -\sum_{j=0}^{T-1}\sum_{i=1}^{n} \mathbf{d}_j^i \log(P^j\mathbf{x}_0)^i & \\ \text{subject to} & P^*\mathbf{1} = \mathbf{1} & \\ & P \geq 0 & \\ & \mathbf{1}^*\mathbf{x}_0 = 1 & \\ & \mathbf{x}_0 \geq 0 & \end{array} \tag{7}$$

Only the objective function is now non-convex, and we can then solve (7) by passing it to the interior point algorithm provided by the fmincon function in MATLAB. As a starting point for P, we use the least squares solution of P generated by solving the constrained least squares problem

$$\begin{array}{lll} \underset{P}{\text{minimize}} & \sum_{j=0}^{T-2} \|P\hat{\mathbf{x}}_j - \hat{\mathbf{x}}_{j+1}\|_2^2 & \\ \text{subject to} & P^*\mathbf{1} = \mathbf{1} & \\ & P \geq 0, & \end{array} \tag{8}$$

where $\hat{\mathbf{x}}_j$ refers to the empirical probability distribution generated by the data at time step j. For the starting point for $\mathbf{x}_0$, we simply use the empirical distribution $\hat{\mathbf{x}}_0$. This direct solution method works very quickly on small problem sizes but scales poorly.

3.2 Alternating Direction Method of Multipliers

We can also solve the full non convex optimization problem (6) by using the alternating direction method of multipliers [Boyd (2010)]. The ADMM approach is slower than the direct minimization but scales better to larger problem sizes, can be parallelized, and has more flexibility for future extensions.

To do ADMM, we add the dynamic constraints into the objective function to form an augmented Lagrangian. The augmented Lagrangian is given by

$$\begin{aligned} \mathcal{L}(P,\mathbf{x},\boldsymbol{\nu}) &= -\ell(\mathbf{x}) \\ &+ \sum_{t=0}^{T-2} \left(\boldsymbol{\nu}_t^* \left(P\mathbf{x}_t - \mathbf{x}_{t+1}\right) + \frac{\rho}{2}\|P\mathbf{x}_t - \mathbf{x}_{t+1}\|_2^2 \right), \end{aligned} \tag{9}$$

where the $\boldsymbol{\nu}$ are Lagrange multipliers. We can then attempt to solve the non convex optimization problem with the iterative ADMM method.

We start by setting the initial values for $\mathbf{x}$ equal to the empirical probability distributions determined by the observations. We also set the Lagrange multipliers $\boldsymbol{\nu}$ to 0 initially. Having defined the zeroth iterates $\mathbf{x}^{(0)}$ and $\boldsymbol{\nu}^{(0)}$, we can proceed with the iterative method.

We take $P^{(k+1)}$ to be the optimal point of the optimization problem

P step:

$$\begin{array}{lll} \underset{P}{\text{minimize}} & \mathcal{L}(P,\mathbf{x}^{(k)},\boldsymbol{\nu}^{(k)}) & \\ \text{subject to} & P^T\mathbf{1} = \mathbf{1} & \\ & P \geq 0, & \end{array} \tag{10}$$

which corresponds to minimizing the augmented Lagrangian (9) over only P with all other variables held constant. We then take $\mathbf{x}^{(k+1)}$ to be the optimal point of the problem

$\mathbf{x}$ step:

$$\begin{array}{ll} \underset{\mathbf{x}}{\text{minimize}} & \mathcal{L}(P^{(k+1)}, \mathbf{x}, \boldsymbol{\nu}^{(k)}) \\ \text{subject to} & \mathbf{1}^T \mathbf{x}_t = 1 \\ & \mathbf{x}_t \geq 0. \end{array} \tag{11}$$

This corresponds to minimizing the augmented Lagrangian over only $\mathbf{x}$.

Having minimized the augmented Lagrangian independently over both P and $\mathbf{x}$, we then perform a dual update step given by

$\boldsymbol{\nu}$ step:

$$\boldsymbol{\nu}_t^{k+1} = \boldsymbol{\nu}_t^k + \rho\,(P\mathbf{x}_t - \mathbf{x}_{t+1}) \tag{12}$$

Both the P step and the $\mathbf{x}$ step can be done efficiently because they are convex optimization problems. For a review of convex optimization, see Boyd and Vandenberghe (2004). The P step is very efficient because it is a quadratic program. The $\mathbf{x}$ step is not as efficient because of the logarithmic terms in the objective. However, the $\mathbf{x}$ step can be parallelized with another layer of ADMM. We use CVX [Grant and Boyd (2008)],[Grant and Boyd (2013)] to perform these convex optimization steps.

The convergence conditions for ADMM consist of checking feasibility of the constraint in the augmented Lagrangian as well as stationarity of the augmented Lagrangian. In our case, we found that the iterates of P and $\mathbf{x}_t$ barely changed after reaching feasibility, and so we terminate once feasibility is reached. Because our problem is not convex, the convergence of ADMM is not guaranteed. However, we typically see convergence in about ten to fifty iterations.

4. IDENTIFIABILITY CONDITIONS

In the previous sections, we discussed the question of how to estimate the transition matrix P from the data. In this section, we consider the question of whether the data uniquely determine P. We also assume that we have infinite sample size, which means at each time step we can measure the true distribution $\mathbf{x}_t$. Our experiments consist of setting $\mathbf{x}_0$ and watching the distribution evolve over time, so we would like to know how to set $\mathbf{x}_0$ so that P is uniquely determined by the time evolution of $\mathbf{x}_t$.

Proposition 1. Let P be a square matrix and $\mathbf{x}_0$ be any initial state. The observations consist of $\mathbf{x}_t = P^t\mathbf{x}_0$ from time $t = 0$ to $t = T - 1$. Then P is uniquely determined by the data if and only if the number of time steps T is greater than the number of states n and $[P - \lambda I\ \mathbf{x}_0]$ has full rank for all real λ.

Proof. We note that P must satisfy $P[\mathbf{x}_0 \dots \mathbf{x}_{T-2}] = [\mathbf{x}_1 \dots \mathbf{x}_{T-1}]$. Each time step imposes n constraints on P and P has n^2 variables, so we need at least n steps, which implies that $T > n$ must hold. Due to Cayley-Hamilton theorem, additional steps past the first n steps will provide redundant information, so we then conclude that $[\mathbf{x}_0 \dots \mathbf{x}_{n-1}]$ must be invertible for the data to determine P uniquely. However, we can rewrite this matrix as $[\mathbf{x}_0 \dots P^{n-1}\mathbf{x}_0]$, which is just the controllability matrix for the linear system $(P, \mathbf{x}_0)$. Thus, we can identify P from the data if and only if $(P, \mathbf{x}_0)$ is controllable. From Dullerud and Paganini (2000), we know that $(P, \mathbf{x}_0)$ is controllable if and only if $[P - \lambda I\ \mathbf{x}_0]$ has full rank for all real λ.

When P is stochastic, the conditions on $\mathbf{x}_0$ are sufficient to determine P from the data but may not be fully necessary if P lies on the boundary of the set of stochastic matrices. However, the takeaway from the proposition is that we must make sure to select $\mathbf{x}_0$ so as to excite all the modes of the system. Also, if the system has eigenspaces with multiple dimension, then there is no single $\mathbf{x}_0$ we can choose to determine P uniquely. Finally, if there are additional affine constraints on P, then these constraints may also render the conditions in the proposition sufficient but not necessary.

5. EXAMPLES

5.1 Two State Example

As a first example, we consider the two state Markov chain given by Figure 1.

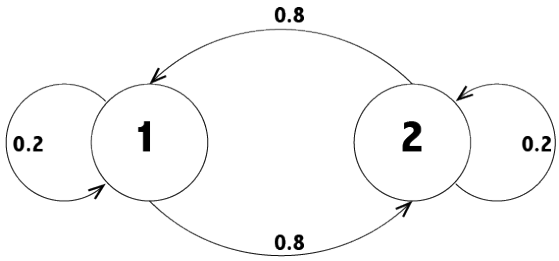

Fig. 1. Two State Example Markov Chain

We then start from an initial distribution of $\mathbf{x}_0 = [0.7\ 0.3]^T$ and simulate the Markov chain for ten steps while collecting twenty samples at each time point. We produce three such time traces and feed the simulated data to both our algorithm as well as the least squares algorithm outlined in (8). We produce two estimates of the probability matrix. P^{ls} is a least squares estimate, and P^{MAP} is the estimate produced by our algorithm. These estimates are given below.

$$\begin{aligned} P^{\text{MAP}} &= \begin{bmatrix} 0.27 & 0.82 \\ 0.73 & 0.18 \end{bmatrix} \\ P^{\text{ls}} &= \begin{bmatrix} 0.40 & 0.67 \\ 0.60 & 0.33 \end{bmatrix} \end{aligned} \tag{13}$$

Our method outperforms the least squares method in terms of producing an estimate of P that is close to the true value. We also ran both these methods on sixty randomly generated traces and kept track of the error distributions. The results are summarized in Figure 2. We see that in general, it is clear that the estimates produced by our method have much less error than those produced by least squares.

In general, our method outperforms least squares in cases where the data is sparse or the chain starts off near the

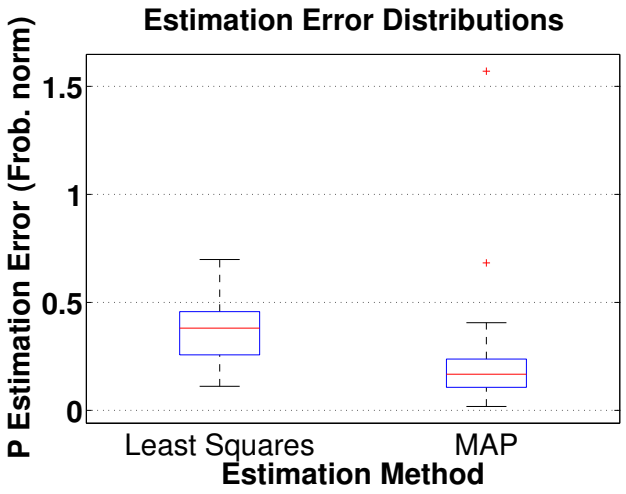

Fig. 2. MAP Estimates Significantly Outperform Least Squares Estimates of P

stationary distribution. In these cases, the fluctuations in measurements become more significant in relation to the dynamics of the chain. Therefore, it becomes more important that we consider these fluctuations as multinomial random variables rather than applying least squares. Also, from the perspective of identifiability, we note that if we started the chain at the uniform stationary distribution of $[0.5\ 0.5]^T$, we would not have been able to solve for P uniquely as any symmetric P has a uniform stationary distribution.

5.2 Toggle Switch

In this section, we apply our method to simulated stochastic gene expression data for a genetic toggle switch [Gardner et al. (2000)]. We simulate the model using the stochastic simulation algorithm (SSA) [Gillespie (1977)] to generate simulated flow cytometry data. We then apply our identification method to the data to attempt to identify the dynamics of the toggle switch.

Our toggle switch model has two different genes and a total of ten species. First of all, we have the DNA for each gene d1 and d2, the RNA for each gene r1 and r2, and the protein for each gene p1 and p2. In addition the proteins can dimerize to form p1p1 and p2p2. Finally, the dimerized proteins can bind to the other gene's DNA to form p2p2d1 and p1p1d2. Copies of the DNA with bound dimers are much less likely to produce RNA, which enforces the mutual repression between the two genes.

The reactions in the system are cataloged in Table 1. The reactions are all symmetric, so only ten of them are shown in the table. The other ten reactions can be formed by switching all the ones and twos in the table and using the same rates. The propensities are proportional to the amount of reactants present as well.

This model provides a very noisy and weakly bistable model of gene expression. In Figure 3, we show an example simulation trace of this model.

We take distributional data on this model by running 100 simulations of the SSA and recording the final amount of

Table 1. Reaction Propensities for Toggle Switch Model

Reaction	Propensity
d1 $\rightarrow$ d1 + r1	0.3
p2p2d1 $\rightarrow$ p2p2d1 + r1	0.002
r1 $\rightarrow$ r1 + p1	2
r1 $\rightarrow \emptyset$	0.5
p1 $\rightarrow \emptyset$	0.067
p1p1 $\rightarrow$ p1	0.067
p1 + p1 $\rightarrow$ p1p1	50
p1p1 $\rightarrow$ p1 + p1	3
p1p1 + d2 $\rightarrow$ p1p1d2	1
p1p1d2 $\rightarrow$ p1p1 + d2	10

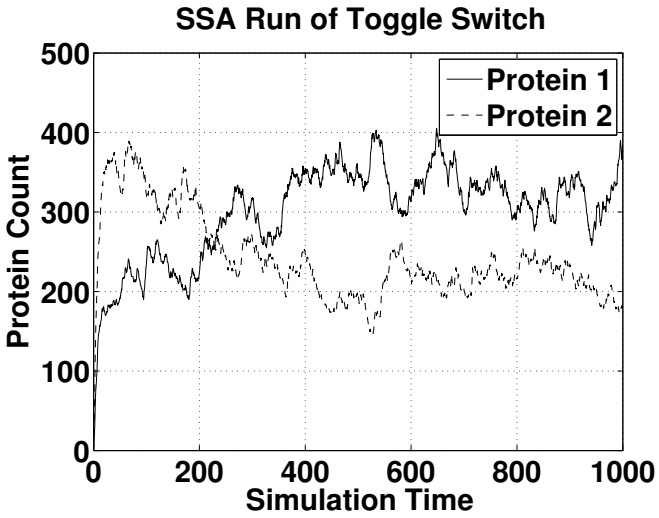

Fig. 3. Example SSA Run

each protein at time intervals of 100 starting from $t = 200$ and going up to $t = 1000$. This gives us 9 time points, with 100 independent trials ending at each time point. We set the initial condition to 50 copies of each DNA, 10 copies of the p2p2 dimer, and 0 copies of everything else. We then run another experiment starting with the same amount of DNA, 10 copies of the p1p1 dimer, and 0 copies of everything else. In an experiment, this type of data could be collected by inducing one gene before data collection. We then discretize each protein into three different levels of expression with cutoffs at 250 copies and 350 copies. This results in a total of 9 discrete states, since each protein can be expressed at a low, medium, or high level. We then bin our measured data and feed it to our algorithm to estimate the transition matrix P, and we enforce that transitions are only allowed to neighboring nodes. The identified model is shown in Figure 4. The graphic excludes transitions with probability less than 0.01, states that are rarely visited, and self transitions.

We see that the model observed is not quite symmetric even though the underlying simulated system is symmetric. However, by looking at the model, we can infer that the system has bistable behavior and tends to spend most of its time with medium to high protein 1 and low protein 2, or medium to high protein 2 and low protein 1. In addition, when the model transitions from one state to the other, it goes through a middle state where both proteins are expressed at a medium level. These insights are consistent with our knowledge of the bistable switch.

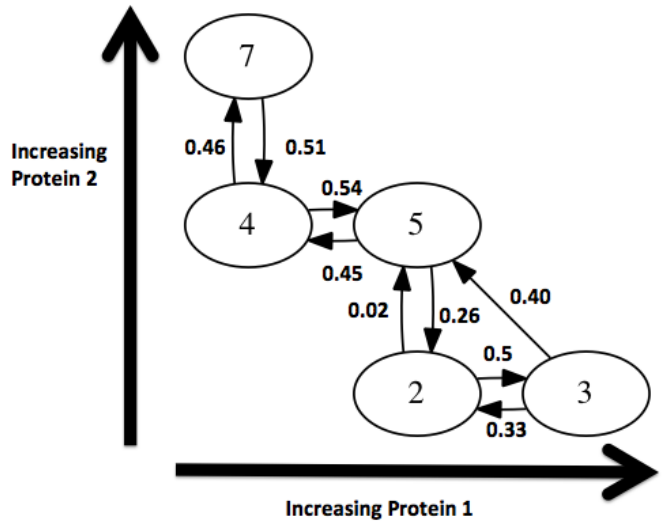

Fig. 4. Identified Toggle Switch Model

The takeaway here is even though the identified model is not very accurate, it provides us with insight into the system that we are studying. Given that the underlying traces are very stochastic and we have a sample size of only 100, it is also hard to expect anything more.

6. DISCUSSION

The method described in this paper is of potential use in a systems biology setting where one might want to study a system of interest in the cell such as a stress response pathway. The key regulators in this pathway may be known, but full knowledge of the system may be unavailable. In that case, this method could be applied to measurements of these key regulators and the identified model could provide insight into how these regulators dynamically interact in order to control the response of the cell.

In the case of the toggle switch example, our method identifies a model that suggests noisy but bistable behavior and also hints that the system usually transitions from one steady state to the other by going through a middle state where both proteins are expressed at a medium level. In addition, we intentionally simulated a small amount of data and made our system highly stochastic. Sparse and highly stochastic data is very common in a biological setting, and we were able to show that at least on a small example, our method can perform adequately in this situation.

Potential future directions of work include working on scalability of the algorithm, automatically selecting a discretization that allows for the best model to be generated, and application of the algorithm to real experimental data.

ACKNOWLEDGEMENTS

We thank Yutaka Hori, Seungil You, Ivan Papusha, and Vipul Singhal for helpful discussions.

REFERENCES

Boyd, S. (2010). Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends in Machine Learning*, 3(1), 1–122.

Boyd, S.P. and Vandenberghe, L. (2004). *Convex optimization*. Cambridge university press.

Dullerud, G.E. and Paganini, F. (2000). *A course in robust control theory*, volume 6. Springer New York.

Eldar, A. and Elowitz, M.B. (2010). Functional roles for noise in genetic circuits. *Nature*, 467(7312), 167–173.

Elowitz, M.B., Levine, A.J., Siggia, E.D., and Swain, P.S. (2002). Stochastic gene expression in a single cell. *Science*, 297(5584), 1183–1186.

Gardner, T.S., Cantor, C.R., and Collins, J.J. (2000). Construction of a genetic toggle switch in escherichia coli. *Nature*, 403(6767), 339342.

Gillespie, D.T. (1977). Exact stochastic simulation of coupled chemical reactions. *The Journal of Physical Chemistry*, 81(25), 2340–2361.

Grant, M. and Boyd, S. (2008). Graph implementations for nonsmooth convex programs. In V. Blondel, S. Boyd, and H. Kimura (eds.), *Recent Advances in Learning and Control*, Lecture Notes in Control and Information Sciences, 95–110. Springer-Verlag Limited.

Grant, M. and Boyd, S. (2013). CVX: Matlab software for disciplined convex programming, version 2.0 beta.

Hori, Y., Khammash, M.H., and Hara, S. (2013). Efficient parameter identification for stochastic biochemical networks using a reduced-order realization. *Proceedings of European Control Conference*, 4154—4159.

Husmeier, D. (2003). Sensitivity and specificity of inferring genetic regulatory interactions from microarray experiments with dynamic bayesian networks. *Bioinformatics*, 19(17), 2271–2282.

Jaynes, E.T. (1957). Information theory and statistical mechanics. *Physical Review*, 106(4), 620–630.

Levin, D.A., Peres, Y., and Wilmer, E.L. (2009). *Markov chains and mixing times*. AMS Bookstore.

Lillacci, G. and Khammash, M. (2011). Model selection in stochastic chemical reaction networks using flow cytometry data. In *Decision and Control and European Control Conference (CDC-ECC), 2011 50th IEEE Conference on*, 16801685.

Lillacci, G. and Khammash, M. (2013). The signal within the noise: efficient inference of stochastic gene regulation models using fluorescence histograms and stochastic simulations. *Bioinformatics*, 29(18), 2311–2319.

Munsky, B. and Khammash, M. (2010). Identification from stochastic cell-to-cell variation: a genetic switch case study. *IET Systems Biology*, 4(6), 356–366.

Neuert, G., Munsky, B., Tan, R.Z., Teytelman, L., Khammash, M., and Oudenaarden, A.v. (2013). Systematic identification of signal-activated stochastic gene regulation. *Science*, 339(6119), 584–587.

Russell, S.J., Norvig, P., Canny, J.F., Malik, J.M., and Edwards, D.D. (1995). *Artificial intelligence: a modern approach*, volume 74. Prentice hall Englewood Cliffs.

Sachs, K., Perez, O., Pe'er, D., Lauffenburger, D.A., and Nolan, G.P. (2005). Causal protein-signaling networks derived from multiparameter single-cell data. *Science*, 308(5721), 523–529.