Probabilistic Approximations of Bio-pathway Dynamics

Bing Liu[1], David Hsu[1,2], P.S. Thiagarajan[1,2]

1 Introduction

Understanding the functioning of complex biological systems is a major challenge. To address this challenge, quantitative mathematical models are needed to capture the dynamics of various intra (and inter)-cellular processes. Here we focus on signaling pathways which constitute the building blocks of many of the intra-cellular processes that govern the behavior of cells.

A standard formalism used to model bio-pathways is a system of ordinary differential equations (ODEs). The equations describe specific bio-chemical reactions, while the variables typically represent concentration levels of molecular species (genes, RNAs, proteins). Bio-pathways usually involve a large number of molecular species and bio-chemical reactions. Hence the corresponding ODE model will involve many variables and parameters and the values of many of the parameters (rate constants) will be unknown. Further, the initial concentration levels of the various species and rate constants will often be available only as *intervals* of values as also the experimental data reporting the measured concentration levels of a small number of proteins at a few time points. In addition, the data will often be gathered using a cell population. Consequently, when numerically simulating the ODEs model, one must resort to Monte Carlo methods to ensure that sufficiently many point values from the relevant intervals of values are being sampled. As a result, tasks such as model validation, parameter estimation and sensitivity analysis will require the generation of a huge number of trajectories.

We propose a probabilistic approach to approximate the deterministic signaling pathway dynamics specified as a system of ODEs. It consists of pre-computing and storing a representative sample of trajectories induced by the system of ODEs. After discretizing the value space suitably, we use Baysian network (BN) models to compactly represent these trajectories by exploiting the dependencies/independecies in the pathway structure. As a result, a variety of analysis questions concerning the pathway dynamics traditionally addressed using Monte Carlo simulations can be converted to Bayesian inference and solved much more efficiently. The BN representation is, in essence, a succinct representation of an associated Markov chain. Hence formal verification techniques developed for Markov chains and other models [2] also become applicable.

2 Methods

We assume the dynamics of a signaling pathway is described by a set of ODEs $\dot{x_i}(t) = f_i(\mathbf{x}(t), \mathbf{k})$ involving the variables $\mathbf{x}$ and parameters $\mathbf{k}$. We then approximate it as a BN. The basic structure of the BN can be derived from the ODEs as illustrated in Figure 1. Each node in BN corresponds to a variable (parameter) in the ODEs at a specific time slice and will have an associated conditional probability table. The parent set of the node x^{t+1} will consist of the node x^t as well as nodes of the form y^t for each y that appear on the right hand side of the ODE corresponding to the variable x. Since the parameter values may be unknown, we often treat them also as variables.

$$S + E \underset{k_2}{\overset{k_1}{\rightleftharpoons}} ES \xrightarrow{k_3} E + P$$

$$\begin{aligned}
\frac{dS}{dt} &= -k_1 \cdot S \cdot E + k_2 \cdot ES \\
\frac{dE}{dt} &= -k_1 \cdot S \cdot E + (k_2 + k_3) \cdot ES \\
\frac{dES}{dt} &= k_1 \cdot S \cdot E - (k_2 + k_3) \cdot ES \\
\frac{dP}{dt} &= k_3 \cdot ES
\end{aligned}$$

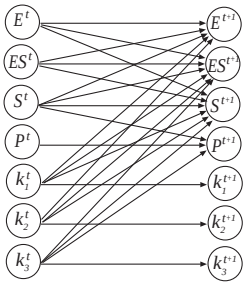

Figure 1: The ODE model of enzyme-kinetic system and its BN representation.

To compute conditional probability tables associated with the nodes, we sample the prior distribution of initial values for the variables and the parameters, and perform numerical integration to generate a sufficiently large number of trajectories. We then discretize those trajectories by the predefined intervals and compute the conditional probabilities for each node by simple counting. This process involves a one time cost and can

[1]NUS Graduate School for Integrative Sciences and Engineering, National University of Singapore. E-mail: lb@nus.edu.sg
[2]Department of Computer Science, National University of Singapore. E-mail: {dyhsu, thiagu}@comp.nus.edu.sg

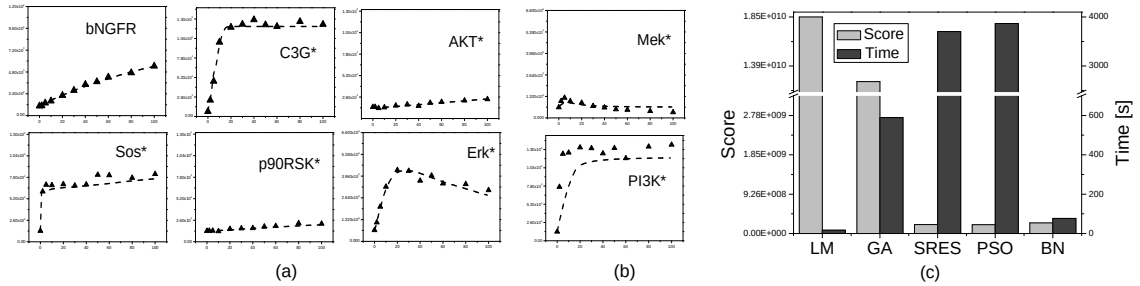

Figure 2: Parameter estimation results. (a) BN simulation profiles vs. training data. (b) BN simulation profiles vs. test data. (c) Performance comparison of our parameter estimation method (BN) and 4 other methods: Levenberg-Marquardt (LM), genetic algorithm (GA), stochastic ranking evolutionary strategy (SRES) and particle swarm optimization (PSO). The scores are the weighted mean squared difference between simulated and experimental data.

be executed in parallel. Analysis tasks can then be performed efficiently using inferencing techniques that are available for Bayesian networks and the one time cost of constructing the BN approximation can be quickly amortized.

For instance, to answer a query such as "what is the concentration of the protein x_i at time t?" one will have to average a representative sample of trajectories that are numerically generated. Using our BN representation, we can provide the answer to this query (and other more sophisticated queries) in terms of marginal probabilities by standard Bayesian inferencing algorithms.

Though each numerical simulation is fast, a huge number of such simulations will be required for performing tasks such as parameter estimation and global sensitivity analysis. In our representation, a single sweep of the inference algorithm will provide information about ensemble of the trajectories encoded by the BN. As a result, the same tasks can be carried out much faster while matching the (lack of) accuracy of experimental data.

3 Results and Discussion

We applied our method on a model of EGF-NGF signaling pathway which contains 32 species and 48 rate constants [1]. After constructing the BN model (5 intervals of values for each variable, 3×10^6 trajectories), we implemented Factored Frontier (FF) algorithm [4] to perform inference on the BN. The resulting time profiles fit the nominal simulation profiles generated by Monte Carlo integration quite well. In terms of running time, a single execution of FF inference is roughly 1000 times faster than generating a stable nominal profile through numerical simulations. We also synthesized time series data for 7 proteins to test the performance of the BN based parameter estimation method. As the parameter search space is discretized, we implemented a direct search algorithm that aims to minimize the difference between BN-based predictions and (synthetic) experimental data. As shown in Figure 2, the BN-simulation profiles generated using estimated parameters have good matches to (the training data and) the test data. We compared the efficiency and quality of our results with 4 optimization algorithms implemented in the COPASI tool [3]. The results shown in Figure 2 suggest that the performance of our method is superior since it obtains good quality parameter estimates in a much shorter time. We further tested the efficiency of our BN-based global sensitivity analysis procedure. As a result, we have identified 4 critical parameters in signal transduction affecting the systems behavior most directly, which is consistent with previous results [5]. Compared to original Monte Carlo strategy, we reduced running time from 22 hours to 34 minutes.

Thus the preliminary results are promising in terms of both accuracy and efficiency. We plan to apply our method to a variety of pathway models in collaboration with biologists. Finally, the BN representation can be viewed as a succinct representation of an associated finite state Markov chain. It will be interesting to develop formal verification techniques -based on the BN representation- to reason about the behavior of the associated Markov chain which in general will be exponentially larger.

References

[1] K. S. Brown, C. C. Hill, G. A. Calero, K. H. Lee, J. P. Sethna, and R. A. Cerione. The statistical mechanics of complex signaling networks: nerve growth factor signaling. *Physical Biology*, pages 184–195, 2004.

[2] J. Heath, M. Kwiatkowska, G. Norman, D. Parker, and O. Tymchyshyn. Probabilistic model checking of complex biological pathways. *Theoretical Computer Science*, 319(3):239–257, 2008.

[3] S. Hoops, S. Sahle, R. Gauges, C. Lee, J. Pahle, N. Simus, M. Singhal, L. Xu, P. Mendes, and U. Kummer. COPASI - a COmplex PAthway SImulator. *Bioinformatics*, 22(24):3067–3074, 2006.

[4] K. Murphy and Y. Weiss. The factored frontier algorithm for approximate inference in DBNs. In *Proceedings of the 17th Conference in Uncertainty in Artificial Intelligence*, pages 378–385, San Francisco, CA, USA, 2001.

[5] S. D. M. Santos1, P. J. Verveer, and P. I. H. Bastiaens. Growth factor-induced MAPK network topology shapes Erk response determining PC-12 cell fate. *Nature Cell Biology*, 9(3):324–330, 2007.