

# Deriving Network Structure from Discrete Data

Patrick M. Dixon<sup>1</sup>, Radek Erban<sup>2</sup>, and Philip K. Maini<sup>3</sup>

**Short Abstract** — The underlying assumption of most gene regulation networks is that the driving processes are best represented by a system of coupled ordinary differential equations (ODEs) or given by stochastic simulation algorithm (SSA). The discrete nature of most experimental data presents a challenge in determining this system. We present a comparison of methods to reconstruct the network of species interactions from discretized data, based on a generalization of the REVEAL algorithm for reconstructing Boolean networks by Liang et al and the method given by Laubenbacher et al.

**Keywords** — reverse engineering, Boolean networks, gene regulation, discretization

## I. INTRODUCTION

THE details of many gene-regulatory pathways are at present essentially black-box processes: The exact nature of the interactions between DNA, mRNA, proteins and other reactants are difficult to observe directly. There are several approaches to modeling these processes given the limited information available. The simplest strategy in dealing with these systems is to treat the regulatory networks as Boolean networks, with binary variables representing the status of different biochemical species. [1] The progress over time of the system is then treated as following Boolean operations, such as AND, OR, and NOT. This type of model can be generalized beyond binary state variables, although interpretation of transition rules becomes murkier at this point. Alternative methods include systems of coupled differential equations, which necessitate the determination of rate constants and interaction patterns, as well as SSAs as pioneered by Gillespie [2]. Note that an advantage of Boolean type models is the reduced need for rate constants. All of these approaches, however, require information about the network of interactions between elements of the system.

We examine a problem related to the inverse of the above. Specifically, we seek to determine the network of interactions between variables given the system dynamics (e.g. obtained by solving ODEs or SSA). The accuracy to which this is possible gives us some idea of the ease with which it is possible to move between the Boolean and rate equation type models.

<sup>1</sup>All authors affiliated with Centre for Mathematical Biolgy, University of Oxford. E-mail: dixonp@maths.ox.ac.uk

<sup>2</sup>E-mail: erban@maths.ox.ac.uk

<sup>3</sup>E-mail: maini@maths.ox.ac.uk

## II. METHODS

Our goal is to examine the ways in which information about the network architecture can be obtained from the data and compare some of the available techniques under various conditions. Here the Boolean framework holds the advantage that information about rate constants is not required for this task. There are several approaches to dealing with the Boolean/algebraic models, of which two are Liang et al.'s REVEAL algorithm, drawing on Shannon entropy, and Laubenbacher et al.'s techniques drawing on the theory of finite fields [3,4]. We examine a generalization of Liang et al.'s REVEAL algorithm, here expanded to apply to multi-state systems. This algorithm is based on measures of Shannon entropy, and specifically with the relationship between two variables' entropies in the case that one is determined by the other. Given a set of input-output vector pairs, the algorithm determines the inputs to each transition rule by means of calculating joint entropies. We apply the algorithm to discrete values obtained by applying thresholds to systems of ODEs.

Selection of the time step,  $\Delta t$  for discretization is crucial, regardless of the algorithm to be applied. Determination of the appropriate choice of  $\Delta t$  in order to obtain the immediate inputs to each variable requires careful fine tuning, and depends on the nature of the network, as well as the number of states used. If  $\Delta t$  is chosen to be too large, then the system may pass through what would otherwise be more than one discrete state between timesteps. If, on the other hand,  $\Delta t$  is chosen to be too small, the system will falsely appear to be in a steady state, due to insufficient time for values to cross thresholds.

We compare the algorithms under a variety of network conditions. We also examine the relationship between the number of discrete states and optimal timestep size.

## REFERENCES

- [1] Guet, C., Elowitz, M., Hsing, W., Leibler, S. "Combinatorial Synthesis of Genetic Networks," *Science*. **296** 1466-1470, 2002.
- [2] Andrews, S., Arkin, A. "Simulating Cell Biology," *Current Biology*. **16**(14) R523-R527, 2006
- [3] Laubenbacher, R., Stigler, B. "A computational algebra approach to the reverse engineering of gene regulatory networks," *J Theoretical Bio*, **229** 523-537
- [4] Liang, S., Fuhrman, S., Somogyi, R., "REVEAL: A general reverse engineering algorithm for inference of genetic network architecture," *Pacific Symposium on Biocomputing* 3, 18-29 (1998)