Reconstructing Chemical Reaction Networks: Data Mining meets System Identification

Yong Ju Cho, Naren Ramakrishnan, and Yang Cao
Department of Computer Science
Virginia Tech, VA 24061, USA
ycho76@cs.vt.edu, naren@cs.vt.edu, ycao@cs.vt.edu

ABSTRACT

We present an approach to reconstructing chemical reaction networks from time series measurements of the concentrations of the molecules involved. Our solution strategy combines techniques from numerical sensitivity analysis and probabilistic graphical models. By modeling a chemical reaction system as a Markov network (undirected graphical model), we show how systematically probing for sensitivities between molecular species can identify the topology of the network. Given the topology, our approach next uses detailed sensitivity profiles to characterize properties of reactions such as reversibility, enzyme-catalysis, and the precise stoichiometries of the reactants and products. We demonstrate applications to reconstructing key biological systems including the yeast cell cycle. In addition to network reconstruction, our algorithm finds applications in model reduction and model comprehension. We argue that our reconstruction algorithm can serve as an important primitive for data mining in systems biology applications.

Categories and Subject Descriptors: H.2.8 [Database Management]: Database Applications - Data Mining; I.2.6 [Artificial Intelligence]: Learning - Induction

General Terms: Algorithms, Measurement, Experimentation.

Keywords: Systems biology, graphical models, Markov networks, ordinary differential equations, network reconstruction.

1. INTRODUCTION

Algorithms in computational biology and bioinformatics are helping rapidly yield new insights into biological and biochemical processes. While much of today's excitement is focused on analyzing data from high-throughput screens (e.g., microarrays, RNAi assays), significant research is also being conducted in constructing and simulating mathematical models of key biological processes, such as the cell cycle [5], circadian rhythms, and entire signaling pathways [2]. These models capture not only qualitative properties of the underlying process but also quantitative traits as revealed by mutant experiments [16]. As shown in Fig. 1, such mathematical modeling typically begins with a chemical reaction network (CRN), which is then converted to a set of simultaneous ordinary differential equations (ODEs), which are then numerically simulated to yield time series profiles of the participating molecular species. These profiles are then matched with real data and the model is adjusted to account for discrepancies. More sophisticated methods involving bifurcation plots and phase portraits shed further insight into the qualitative dynamics of the underlying system.

In this paper, we study the inverse problem, i.e., analyzing time series profiles of the molecular species to reconstruct the CRN (see Fig. 1, dotted lines). This finds uses in not just systems biology, as studied here, but also in any domain where chemical reaction systems form the origins of the underlying numerical model (ODE), such as petrochemical plant engineering, environmental engineering, food processing, and manufacturing.

Reconstructing CRNs is relevant not just for system identification but also for model reduction. For instance, it is well acknowledged that models of key biological processes are notoriously complex and difficult to comprehend for humans [2]. A key task therefore is to reduce the reaction system to a smaller system, involving fewer reactions and/or molecules, but yet retain the essential dynamical properties of the system. Given a complex mathematical model of, say, a biochemical process, we can simulate the model to generate data and reconstruct a (potentially) smaller model by mining the generated dataset. Such a model → data → model transformation is currently a hot topic in computational systems biology [14].

Pertinent data for mining CRNs can hence be gathered from either experimental observations or computational simulation. The former is the subject of works such as [15] and requires 'wet-lab' machinery as described in [1]. In this paper, we focus on data from computational simulations of mathematical models for three reasons: the ease of generating data on demand from the given CRN in a controlled fashion, the capability to systematically perturb the CRN and observe the modified dynamics, and the desire to verify our algorithms on some 'ground truth.' Table 1 summarizes the input-output description of the network reconstruction problem studied here as well as the methods available to observe, interrupt, or otherwise modify the behavior of the system. This setting of the CRN mining problem is pertinent in computational modeling and systems biology contexts.

KDD'08, August 24–27, 2008, Las Vegas, Nevada, USA.
Copyright 2008 ACM 978-1-60558-193-4/08/08 ...$5.00.

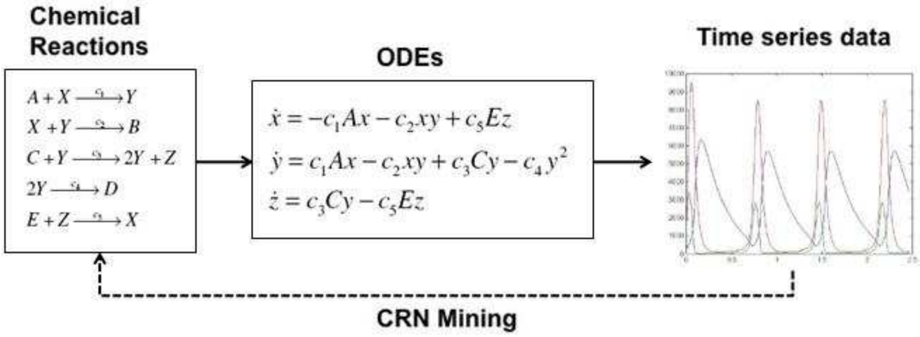

Chemical Reactions
$A+X \xrightarrow{c_1} Y$
$X+Y \xrightarrow{c_2} B$
$C+Y \xrightarrow{c_3} 2Y+Z$
$2Y \xrightarrow{c_4} D$
$E+Z \xrightarrow{c_5} X$
ODEs
$\dot{x} = -c_1Ax - c_2xy + c_5Ez$
$\dot{y} = c_1Ax - c_2xy + c_3Cy - c_4y^2$
$\dot{z} = c_3Cy - c_5Ez$
Time series data
CRN Mining

Figure 1: CRN mining is the inverse problem of reverse-engineering a set of chemical reactions that can reproduce the dynamics observed in a given time series dataset.

Our primary contributions in this paper are four fold. First, we introduce CRN mining as a new KDD problem and cast CRN mining as the task of mining an undirected graphical model followed by annotating edges and groups of edges with chemical reaction type information. In essence, we capture the dynamics of the network by modeling each species as a random variable and by looking for independence relations between them.

A key issue in mining graphical models among a given set of random variables is to decide whether to detect dependencies or (conditional) independencies. If we choose to detect dependencies, we must take care to distinguish between direct and indirect dependencies. To avoid this issue, classical algorithms (e.g., see [3]) are hence almost exclusively based on detecting independencies, either by explicitly identifying such constraints and summarizing them into a network, or by defining the score of a network based on such relationships and searching in the space of networks. Our second contribution is to show how the novel setting of CRN mining permits us to mine dependencies and yet avoid detecting indirect dependencies, a feature not achievable in traditional (discrete) graphical model mining contexts. Further, our algorithm for CRN mining involves a $O(n^2)$ computation (where n is the number of species) in contrast to algorithms that have exponential running time complexity in the worst case for mining graphical models.

Our third contribution is the notion of 'sensitivity tables' as pattern matching constraints to identify reaction types, such as whether it is a reversible or irreversible reaction, enzyme catalyzed or not, and the precise ratios between the molecules of reactants and products. We hasten to add that we cannot unambiguously distinguish between all possible chemical reaction types and we precisely state the distinctions that we are (un)able to make.

Finally, we demonstrate the application of CRN mining to reconstructing many important biochemical networks in systems biology applications, including prokaryotic gene expression regulation and the CDC-Cyclin2 interaction forming the core of the budding yeast cell cycle.

2. RELATED RESEARCH

CRNs have been well studied in bioinformatics applications. Most of the dynamic behavior of cells can be reduced to the underlying (bio)chemistry of how molecules such as genes, proteins, and RNA interact, catalyze reactions, and contribute to the proper functioning of cells. Hence studying a biological system by casting it as a CRN is typically the first step in mathematical modeling. For our purposes here, we focus on research that attempts to reconstruct CRNs.

Table 1: Setting of the CRN mining problem.

Given
Number of species
Identities of species
Time series profiles of molecular concentrations

To find
Reaction network
Properties of individual reactions

Perturbation capabilities
Can buffer given species (either singly or in subsets)
Can knock-out given species (either singly or in subsets)

The 1997 paper by Arkin, Shen, and Ross in *Science* [1] is credited with creating interest in CRN mining; it also presented an all-pairs correlation method for reconstructing the underlying network, with applications to the glycolysis metabolic process. However, the method described in [1] cannot distinguish between direct and indirect dependencies and can thus result in spurious edges. In addition, it assumes that all species are eventually connected and hence cannot recognize disconnected components, such as the simultaneous set of chemical reactions: $\{A \longleftrightarrow B, C \longleftrightarrow D\}$.

There have been many papers that were motivated by the Arkin, Shen, and Ross work described above. For instance, Wiggins and Nemenman [19] present a method to analyze time series to infer process pathway, which can be construed as representing calling invocations of one pathway by another. However, their method is aimed at producing a general network of relationships from genomic data and not at reconstructing chemical reaction networks. A more theoretical approach is taken in [13] but its strong guarantees of the soundness of network reconstruction are obtained by restricting the focus to discrete dynamical systems, which capture the functional behavior of regulatory networks but not CRNs. More recently, Karnaukhov et al. [9] focus on the reaction identification problem by assuming a general parameterized form for the kinetics of the reaction and fitting rate constants by least squares fitting. This work builds on earlier work by the same authors [8]. CRN mining as studied here subsumes reaction identification as a sub-goal.

Thus, our formulation of CRN mining is novel for its attempt to model both the dependence structure of chemical species and the properties of individual reactions.

3. SOME CHEMISTRY FOR DATA MINERS

Before we present our algorithm for reconstructing chemical reaction networks, we review some basic chemistry and established practices in the mathematical modeling of chemical reactions. This is the subject of many excellent books, such as [11] which especially focus on modeling for bioinformatics applications. For the data mining audience, we present an abridged version of this literature involving only topics necessary to understand the ensuing algorithm.

3.1 Modeling a Single Reaction

The simplest example of a chemical reaction is the irreversible isomerization reaction

$$A \xrightarrow{k_1} B. \tag{1}$$

where k_1 denotes the rate at which species A is converted into B. If the concentrations of the species A and B are represented by x_A and x_B, the dynamics of (1) can be formulated by a set of ordinary differential equations (ODEs)

$$\left\{ \begin{array}{rcl} \frac{dx_A}{dt} &=& -k_1 x_A, \\ \frac{dx_B}{dt} &=& k_1 x_A. \end{array} \right. \tag{2}$$

A typical trajectory of x_A and x_B in this simple system is shown in Figure 2 (a).

The reaction (1) is a special case of the reversible isomerization reactions

$$A \underset{k_2}{\overset{k_1}{\longleftrightarrow}} B. \tag{3}$$

The corresponding ODEs are:

$$\left\{ \begin{array}{rcl} \frac{dx_A}{dt} &=& -k_1 x_A + k_2 x_B, \\ \frac{dx_B}{dt} &=& k_1 x_A - k_2 x_B. \end{array} \right. \tag{4}$$

A typical trajectory for this system is shown in Figure 2 (b).

Both reactions (1) and (3) are linear. The simplest nonlinear example is the bimolecular reaction

$$A + B \xrightarrow{k_1} C. \tag{5}$$

The corresponding ODEs are given below.

$$\left\{ \begin{array}{rcl} \frac{dx_A}{dt} &=& -k_1 x_A x_B, \\ \frac{dx_B}{dt} &=& -k_1 x_A x_B, \\ \frac{dx_C}{dt} &=& k_1 x_A x_B. \end{array} \right. \tag{6}$$

A typical trajectory of equation (6) is shown in Figure 2 (c).

The kinetics in reactions (1), (3) and (5) are simple mass action kinetic laws. But equations can be more complicated. Consider the enzyme-substrate reactions

$$E + S \underset{k_{-1}}{\overset{k_1}{\longleftrightarrow}} ES \xrightarrow{k_2} E + P. \tag{7}$$

Here E represents enzyme species, whose total concentration $E_0 = x_E + x_{ES}$ remains as a constant in this chemical process. The corresponding ODEs are

$$\left\{ \begin{array}{rcl} \frac{dx_S}{dt} &=& -k_1 x_E x_S + k_{-1} x_{ES}, \\ \frac{dx_E}{dt} &=& -k_1 x_E x_S + (k_{-1} + k_2) x_{ES}, \\ \frac{dx_{ES}}{dt} &=& k_1 x_E x_S - (k_{-1} + k_2) x_{ES}, \\ \frac{dx_P}{dt} &=& k_2 x_{ES}. \end{array} \right. \tag{8}$$

When k_1 and k_{-1} are much larger than k_2, we can assume the first two reactions in (7) reach partial equilibrium. This partial equilibrium assumption can be formulated by

$$k_1 x_E x_S = k_{-1} x_{ES}. \tag{9}$$

When k_2 is in a similar magnitude of k_{-1}, the equilibrium assumption (9) does not hold any more. But a steady state assumption can be made. It assumes that the concentration of ES remains a steady state after a transient period, which is formulated as

$$k_1 x_E x_S = (k_{-1} + k_2) x_{ES}. \tag{10}$$

It turns out that (9) is a special case of (10). Let $k_M = \frac{k_2 + k_{-1}}{k_1}$. With the assumption that E_0 is much smaller than x_S, we can derive

$$\frac{dx_P}{dt} = \frac{k_2 E_0}{k_M + x_S} x_S. \tag{11}$$

Let $k = \frac{k_2 E_0}{k_M + x_S}$. The equation (11) is called the Michaelis-Menten equation. It reduces the enzyme-substrate reaction (7) into a simple reaction

$$S \xrightarrow{E} P. \tag{12}$$

denoting that substrate S is catalyzed by enzyme E to form product P. But (12) is fundamentally different from the simple reaction (1) because it follows the nonlinear enzyme kinetics (11). A typical trajectory of the reaction (12) is shown in Figure 2 (d).

3.2 Modeling Sets of Reactions

A chemical reaction network (CRN) is composed of many reactions. Suppose N species are involved in M reaction channels in a CRN. Let the concentration of these species be denoted by x_i, $i = 1, \cdots, N$ and the reaction channels be denoted by R_j, $j = 1, \cdots, M$. The dynamics of the system can be formulated as

$$\frac{dx}{dt} = f(x), \tag{13}$$

where $f_i(x) = \sum_{j=1}^{M} \nu_{ij} r_j(x)$. Here ν is called the stoichiometric matrix. ν_{ij} is the unit change of x_i caused by the reaction channel R_j and $r_j(x)$ is the reaction rate function for the reaction channel R_j. For example, in the simple reaction (1), there are two species and one reaction channel. $\nu = [-1,\ 1]$ and $r_1(x) = k_1 x_A$. In the bimolecular reaction (5), $\nu = [-1,\ -1,\ 1]$ and $r_1(x) = k_1 x_A x_B$. In the reduced enzyme-substrate reaction (12), $\nu = [-1,\ 1]$ and $r_1(x) = \frac{k_2 E_0}{k_M + x_S}$.

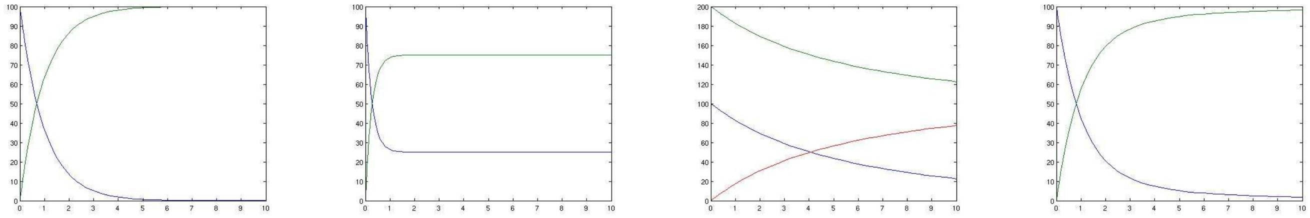
But often the state space in (13) can be reduced by applying conservation laws and partial equilibrium or steady state assumptions. Examples of the partial equilibrium assumption and steady state assumption are given in (9) and (10) for the enzyme-substrate reaction (7). Conservation laws can be applied for all examples shown above. For example, for reaction systems (1) and (3), the sum of x_A and x_B remains as a constant. That can be formulated as

$$x_A + x_B = C_0. \tag{14}$$

With this conservation law, we only need to formulate the dynamics of one variable. The other can be directly calculated from (14). Thus the dimension of the state space in both equations (2) and (4) can be reduced by 1. In the bimolecular reaction (5), there are two conservation laws

$$\left\{ \begin{array}{rcl} x_A + x_C &=& C_0, \\ x_B + x_C &=& C_1. \end{array} \right. \tag{15}$$

With the two constraints, the dimension of the state space in equation (6) can be reduced to 1.



(a) $A \xrightarrow{k_1} B$ (b) $A \underset{k_2}{\overset{k_1}{\rightleftarrows}} B$ (c) $A + B \xrightarrow{k_1} C$ (d) $S \xrightarrow{E} P$

Figure 2: Dynamics of reactions 1, 3, 5, and 12, respectively. Parameters used in the above plots: (a) $k_1 = 1$, $x_A(0) = 100$ **and** $x_B(0) = 0$. **(b)** $k_1 = 3$, $k_2 = 1$, $x_A(0) = 100$ **and** $x_B(0) = 0$. **(c)** $k_1 = 0.001$, $x_A(0) = 100$, $x_B(0) = 200$ **and** $x_C(0) = 0$. **(d)** $k_1 = 1$, $k_{-1} = 10$, $k_2 = 1$, $x_S(0) = 100$ **and** $x_P(0) = 0$.

For a complex CRN, the ODEs and the algebraic constraints can be put together. Then we obtain a set of differential-algebraic equations (DAEs)

$$x' = f(x, y), \tag{16}$$
$$0 = g(x, y), \tag{17}$$

where (16) is the differential part and (17) is the algebraic part.

3.3 Sensitivity Analysis

Sensitivity analysis is widely used in optimization, parameter estimation, uncertainty and stability analysis. (Here we demonstrate its applications to data mining and network reconstruction.) For a CRN represented by a set of DAEs, the system often contains uncertainty due to unknown kinetic rates, environment fluctuations, and other unknown possible reaction pathways. They can be represented as parameters in DAEs. We can rewrite the equation (16-17) as

$$x' = f(x, y, p), \tag{18}$$
$$0 = g(x, y, p), \tag{19}$$

with initial conditions $x_0 = x_0(p)$ and $y_0 = y_0(p)$. Sensitivity reflects the change rates of the state variables x and y with respect to the change in the parameter p, which are calculated by $\frac{dx}{dp}$ and $\frac{dy}{dp}$.

The sensitivity functions $\frac{dx}{dp}(t)$ and $\frac{dy}{dp}(t)$ can be obtained from the numerical time series data or estimated by finite difference methods during the process of solving the original DAEs and derived sensitivity equations. Software such as DASPK (in Fortran) [4] and CVODES (which comprises the CVODE [6], KINSOL, and IDE software components in C) have in-built capabilities to perform sensitivity analysis of DAEs.

4. USING SYSTEMATIC PROBING TO IDENTIFY CRNS

Referring back to the experimental context in Table 1, we present an approach to reconstructing chemical reaction networks by systematically perturbing the network to identify relationships between the given species. (Although such perturbations are well studied in biochemistry, leading to the notion of minimal cut sets in biochemical networks [10], they have primarily been used for engineering flux patterns, not for CRN mining.) As Table 1 shows, there are two main classes of perturbations available: buffering and knock-out experiments.

4.1 Buffering experiments

Buffering involves providing enough supply (intake) of some species, thus forcing it to stay constant. In the corresponding DAEs, this is equivalent to replace the corresponding differential equation by a simple algebraic equation. Note that buffering will break the corresponding conservation constraints.

For example, consider a simple chain reaction system

$$A \xrightarrow{k_1} B \xrightarrow{k_2} C. \tag{20}$$

The corresponding equations are

$$\begin{cases} \frac{dx_A}{dt} &= -k_1 x_A, \\ \frac{dx_C}{dt} &= k_2 x_B, \\ x_A + x_B + x_C &= C_0. \end{cases} \tag{21}$$

If we perturb the initial value of A (let $x_A(0) = p$), we can calculate the corresponding change resulted in C (by $\frac{dx_C}{dp}$). We then know A and C are connected in the system. If B is buffered, x_B stays as a constant. Then the equations become

$$\begin{cases} \frac{dx_A}{dt} &= -k_1 x_A, \\ \frac{dx_C}{dt} &= k_2 x_B, \\ x_B &= B_0. \end{cases} \tag{22}$$

We conduct the sensitivity analysis again and we will get $\frac{dx_C}{dp} = 0$! This shows that after B is buffered, A and C become disconnected. We can then conclude about the structure of this network: A affects C through B.

4.2 Knock-out experiments

A second type of perturbation that is common in biology is the knock-out, i.e., to remove a molecule completely by rendering it inactive or unable to participate in the reaction. Engineered biological systems by knocking out key molecules are referred to as *mutants*. In the corresponding DAE, knock-outs correspond to a special form of buffering, namely replacing the respective species variables to zero.

However, knock-outs, while useful at understanding loss-of-function, are not very revealing for reconstructing CRNs. For instance, compare the chain reaction:

$$A \longrightarrow B \longrightarrow C$$

with the enzyme catalyzed reaction:

$$A \xrightarrow{B} C$$

By buffering B, we can distinguish between the two cases by detecting whether $\frac{dx_C}{dp} = 0$ (first case) or whether $\frac{dx_C}{dp} > 0$ (second case). Here p is the initial value of A as before. However, if we knock out B from the respective equations, both of them result in $\frac{dx_C}{dp} = 0$! For this reason, in this paper, we exclusively focus on buffering as a means to probe CRNs.

4.3 CRNs and Graphical Models

The above observations hint at the relationship between CRNs and undirected graphical models [12]. We first setup the correspondence between a given CRN and a corresponding graphical model. For ease of presentation, in the following lemmas and results, we assume only bimolecular reactions (i.e., each reaction connects only two species) although our algorithmic implementation and experimental results involve both bimolecular and trimolecular reactions.

Definition 1. *Given a CRN $\mathcal{N}$ (a set of molecular species and a set of chemical reactions between them) we define the undirected graph $\mathcal{G}(\mathcal{N})$ corresponding to $\mathcal{N}$ as the graph whose nodes corresponds to the species in $\mathcal{N}$ and whose edges connect nodes that participate in a common reaction.*

Note that different CRNs might induce the same undirected graphical model. For instance, the reaction sets $A \longleftrightarrow B \longleftrightarrow C$ and $A \longrightarrow B \longrightarrow C$ induce the same graph even though the former involves reversible reactions and the latter involves irreversible reactions. Nevertheless, the following results (stated without proof due to space limitations) demonstrate that mining graphical models is an useful first step to reconstructing CRNs.

Lemma 4.1. *Given a network $\mathcal{N}$ and its undirected graph $\mathcal{G}(\mathcal{N})$, node n_1 is conditionally independent of node n_2 given a set of nodes n_X in $\mathcal{G}(\mathcal{N})$ iff the following applies: after buffering n_X in $\mathcal{N}$, the sensitivity of n_1 to n_2 (and vice versa) is zero.*

A direct application of Lemma 4.1 would require us to search through an exponential set of possible conditioning contexts. Instead, as stated earlier, we will seek to identify dependencies.

Lemma 4.2. *Given a network $\mathcal{N}$ and its undirected graph $\mathcal{G}(\mathcal{N})$, an edge exists between node n_1 and node n_2 in $\mathcal{G}(\mathcal{N})$ iff the following applies: the sensitivity of n_1 to n_2 (or vice versa) after buffering all other molecules in $\mathcal{N}$ is non-zero.*

Unlike Lemma 4.1, Lemma 4.2 requires only a search through $O(n^2)$ conditioning contexts. Then why don't traditional Markov network learning algorithms utilize a similar approach? This is because to verify each of the $O(n^2)$ conditional dependencies, the conditioning set involve $n-2$ variables and, even if each variable takes on only two values, we will have to investigate 2^{n-2} settings for conditioning contexts. Besides the exponential complexity, projecting to $n-2$ variables typically will retain very few tuples, typically not sufficient to estimate dependence. Other works such as [3] acknowledge these issues and, in fact, incorporate the size of the conditioning context in their analysis of algorithm complexity. However, in CRN mining, these limitations do not apply since there is a proportional, rather than exponential, cost to a buffering experiment w.r.t. the size of the conditioning context (i.e., the number of buffered molecules). Furthermore, the limitations of sample data sizes do not obviously arise in a buffering experiment.

5. ALGORITHMS FOR CHEMICAL REACTION NETWORK RECONSTRUCTION

Our approach to CRN reconstruction begins by first reconstructing the underlying graphical model (Algorithm 1: InferGraphicalModel) followed by cataloging the individual edges or groups of edges into reactions (Algorithm 2: FindReactions). These are detailed next.

Algorithm 1 InferGraphicalModel

```
Input: $V, ODE_v$
Output: $S$
  for all $i, j \in V (i < j)$ do
    $(S(i,j), S(j,i)) \leftarrow$ BufferedSim$(i, j, V - \{i,j\}, ODE_V)$
  end for
```

Algorithm 2 FindReactions

```
Input: $V, S$
Output: $Bi, Tri$
  for all $i, j \in V (i < j)$ do
    if $|S(i,j)| \geq stol$ or $|S(j,i)| \geq stol$ then
      $E \leftarrow E \cup \{i,j\}$
    end if
  end for
  Initialize all elements of $CV$ to be 0
  $SI \leftarrow sign(S, stol)$
  for all $e_k, e_m \in E (k < m)$ do
    if $e_k$ and $e_m$ share a vertex $b$ s.t. $e_k = \{a,b\}$ and
    $e_m = \{b,c\}$ and $Tri$.find$(\{a,b,c\})$ = false then
      reactions $\leftarrow$ LookupTriReaction$(\{a,b,c\}, SI)$
      if reactions is not empty then
        $Tri$.add$(\{a,b,c\}$, reactions)
        set $CV(\{a,b\}), CV(\{b,c\}), CV(\{c,a\})$ to be 1
      end if
    end if
  end for
  for all $e = \{h,i\} \in E$ do
    if $CV(\{h,i\}) = 0$ then
      reactions $\leftarrow$ LookupBiReaction$(\{h,i\}, SI)$
      if reactions is not empty then
        $Bi$.add$(\{h,i\}$, reactions)
      end if
    end if
  end for
```

5.1 Reconstructing Network Topology

InferGraphicalModel takes as input V, the set of all chemical species whose dynamics are given by the system of ODEs in ODE_V. As stated earlier, it conducts a $O(n^2)$ buffered simulation to identify sensitivities between all pairs of molecules (in both directions). Here, $S(i,j)$ denotes the sensitivity of j to the initial concentration of i. InferGraphicalModel produces as output the sensitivity matrix S whose non-zero entries encode the graphical model.

The next algorithm, FindReactions, takes as input the set of chemical species as before and the just computed sensitivity matrix S. It produces as output the list of detected bimolecular reactions in Bi and trimolecular reactions in Tri. First, it thresholds the sensitivity matrix S into SI. The array CV is used to hold a CoVer for the molecular species and their dependencies, i.e., to see if a dependency detected in InferGraphicalModel has been ‘explained’ by a chemical reaction. Initially no dependencies are explained, hence CV, indexed by the dependencies, is initialized to zero. Algorithm FindReactions then proceeds to look for trimolecular reactions that fit the sensitivity profiles computed in SI (using Table 3, explained in the next section) and if a suitable reaction is found, the array CV is updated suitably. Only after all trimolecular combinations are exhausted does it proceed to look for bimolecular reactions. At this point, it is important to mention that the algorithm LookupTriReaction (not detailed here) searches through all permutations of the given triple of molecules in establishing a correspondence to sensitivity profiles.

5.2 Reconstructing Reaction Properties

It remains to be detailed how LookupTriReaction and LookupBiReaction work. The advantage to these algorithms is that they use sensitivities between pairs of molecules which can actually be computed alongside the reconstruction algorithm. Tables 2 and 3 contain the relevant information for disambiguating reaction types. The same information is also summarized graphically in Fig. 3. Rather than go through each entry sequentially, we explain below how the sensitivity table patterns can be used to make important distinctions.

Sensitivity changes with time. Let $s_{A,B}(t)$ be the time series of sensitivity of B to the initial concentration of A. We first discretize this time series into ‘+’, ‘-’, and 0 values. The sign of the sensitivity profile, $s(A, B)$, is then defined as the sign of $s_{A,B}(t_i)$ where t_i is the time point at which $|s_{A,B}(t_i)|$ is maximum. We index into Tables 2 and 3 using these signs and identify reaction types. Recall that Table 2 is meant to be used for identifying reactions between pairs of molecules *after* Table 3 has been used to identify reactions between triples. Also, Table 3 is richer in detail than Table 2 since it gives the signs of sensitivities of six basic trimolecular reactions: $A \xrightarrow{B} C$, $A \longleftrightarrow B + C$, $A \longrightarrow B + C$, $A \xrightarrow{A} B + C$, $A \xrightarrow{B} B + C$, and $A + B \longrightarrow C$, and under three different buffering conditions.

We should point out that not all distinctions can be made unambiguously. For instance, in Table 2[1], there are five possible reactions but only three distinct sensitivity patterns. Hence some rows lead to multiple hypotheses. A direction of future work is to develop a constraint engine that can reason about such multiple hypotheses, across adjacent sensitivity profiles, to achieve greater discrimination of detection.

5.2.1 *Reversible versus Irreversible*

Distinguishing between reversible and irreversible reactions is straightforward, e.g., Table 2 can be readily used to distinguish between $A \longrightarrow B$ and $A \longleftrightarrow B$ by assessing the sign of $s(B, A)$.

5.2.2 *Multiple reactants*

This situation requires us to distinguish between the trimolecular reaction $A + B \longrightarrow C$ and the combined set of two bimolecular reactions $\{A \longrightarrow C,\ B \longrightarrow C\}$. $s(A, B)$ and $s(B, A)$ are zero for the two bimolecular reactions but $s(A, B)$ and $s(B, A)$ are negative in the trimolecular reaction, thus enabling the distinction.

5.2.3 *Multiple products*

This situation is the converse of the previous case. Note that $A \longrightarrow B + C$ and the combined set of two bimolecular reactions $\{A \longrightarrow B,\ A \longrightarrow C\}$ have the same signs of sensitivities according to Tables 2 and 3. Thus, $A \longrightarrow B + C$ and $\{A \longrightarrow B,\ A \longrightarrow C\}$ cannot be distinguished in our approach.

5.2.4 *Stoichiometry*

Stoichiometry refers to the relative ratios of molecules that participate in a reaction. Thus, the only distinction between the reactions: $A \longleftrightarrow B$ and $2A \longleftrightarrow B$ is one of stoichiometry. Using only the signs of the sensitivity entries, these reactions cannot be disambiguated. On the other hand, if information about the magnitude of the sensitivity is available, e.g., if we know that $\frac{s_{A,A}(t)}{s_{B,A}(t)} \approx c$ and $\frac{s_{A,B}(t)}{s_{B,B}(t)} \approx c$, then we can conclude the existence of reaction $cA \longleftrightarrow B$ in steady state.

5.2.5 *Enzyme catalysis*

An enzyme-substrate reaction can be modeled with either mass action kinetics or Michaelis-Menten kinetics. When the enzyme-substrate reaction is modeled with mass action kinetics, the sensitivity profiles are identical for $A \xrightarrow{B} C$ and $A + B \longrightarrow C$ (see row 3 of Table 3). On the other hand, if the enzyme-substrate reaction is modeled with Michaelis-Menten kinetics, then these reactions can be disambiguated (see row 4 of Table 3).

5.2.6 *Auto-catalysis*

Auto-catalysis is the situation where a molecule catalyzes a reaction that it itself participates in. It is easier to detect if the catalyst is the product, rather than the reactant. For instance, as can be seen in Table 2, $A \longrightarrow B$ and $A \xrightarrow{A} B$ have the same sensitivity profile, whereas $A \longrightarrow B$ and $A \xrightarrow{B} B$ can be distinguished. Similarly, in Table 3, $A \longrightarrow B + C$ and $A \xrightarrow{A} B + C$ have the same sensitivity profile (see row 2) and thus cannot be distinguished.

5.2.7 *Detecting Groups of Reactions*

The last two rows of Table 3 are especially designed to detect common groups of reactions. The ‘+’ sign for $s(C, A)$ in both these rows helps detect the existence of a loop back from molecule C to A which is not the case, for instance, in rows 3 and 4 of Table 3. Within the last two rows, further disambiguation about rate laws can be made using the sign of $s(A, B)$.

5.2.8 *More Complex Dynamics*

By capturing more of the dynamics, these tables can be put to further use in reaction identification. For instance, consider the task of distinguishing $A \xrightarrow{B} C$ from $A + B \longrightarrow C$ (using rows 3 and 6 of Table 3). When A is buffered, $s(A, C)$ and $s(B, C)$ grow boundlessly in $A \xrightarrow{B} C$. Whereas, in $A + B \longrightarrow C$, $s(A, C)$ is limited by B. Hence, $s(A, C)$ stops increasing after reaching steady state.

[1] A note about the asterisk in this table: due to the process of enzyme-substrate complex formation, the entry $s(B, A)$ is negative for the initial reaction and later changes its sign to a plus as shown in Table 2. If we assume that the (initial) concentration of B is much smaller than the concentration of A, then this entry can be treated as a ‘+’.

Table 3: The 'All but 2' sensitivity table used to identify chemical reactions involving 3 molecules.

Reaction(s)	A buffered		B buffered		C buffered	
	s(B,C)	s(C,B)	s(A,C)	s(C,A)	s(A,B)	s(B,A)
$A \longleftrightarrow B + C$	-	-	+	+	+	+
$A \longrightarrow B + C$ or $A \xrightarrow{A} B + C$	0	0	+	0	+	0
$A \xrightarrow{B} C$ or $A + B \longrightarrow C$	+	0	+	0	-	-
$A \xrightarrow{B} C$ (Michaelis-Menten)	+	0	+	0	0	-
$A \xrightarrow{B} B + C$	+	0	+	0	+	-
$A \xrightarrow{B} C$ or $A + B \longrightarrow C$ with $C \longrightarrow A$	+	0	+	+	-	-
$A \xrightarrow{B} C$ with $C \longrightarrow A$ (Michaelis-Menten)	+	0	+	+	0	-

Table 2: The Bimolecular sensitivity table used to identify chemical reactions involving 2 molecules.

Reaction	s(A,B)	s(B,A)
$A \longrightarrow B$ or $A \xrightarrow{A} B$	+	0
$A \longleftrightarrow B$ or $2A \longleftrightarrow B$	+	+
$A \xrightarrow{B} B$	$+^*$	-

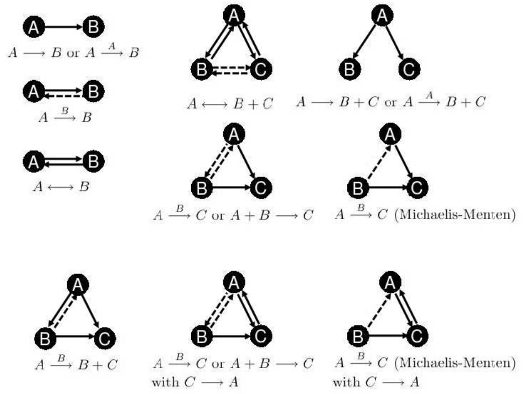

A
B
$A \longrightarrow B$ or $A \xrightarrow{A} B$
$A \xrightarrow{B} B$
$A \longleftrightarrow B$
C
$A \longleftrightarrow B + C$
$A \longrightarrow B + C$ or $A \xrightarrow{A} B + C$
$A \xrightarrow{B} C$ or $A + B \longrightarrow C$
$A \xrightarrow{B} C$ (Michaelis-Menten)
$A \xrightarrow{B} B + C$
$A \xrightarrow{B} C$ or $A + B \longrightarrow C$ with $C \longrightarrow A$
$A \xrightarrow{B} C$ (Michaelis-Menten) with $C \longrightarrow A$

Figure 3: A graphical notation (not meant to be a probabilistic graphical model) of the information from Tables 2 and 3. A solid arrow from node X to node Y exists if sensitivity of Y to initial value of X is positive. A dashed arrow from node X to node Y exists if sensitivity of Y to initial value of X is negative. No arrow denotes a sensitivity of zero.

6. LIMITATIONS AND POSSIBLE SOLUTIONS

Thus far, we have made two critical assumptions that are necessary to the success of our reconstruction algorithm:

1. Between a given pair or triple of molecules, there is at most one reaction.

2. The rate laws governing the reactions fall into the categories of either the mass-action formulation (equations 2) or Michaelis-Menten kinetics (equation 11).

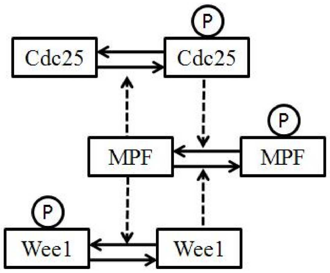


Figure 4: CRN governing cell-cycle transitions in frog egg extracts.

These assumptions are not difficult to surmount but their removal is beyond the scope of this paper. Consider for instance the network in Fig. 4 governing how cells in frog egg extracts divide. The core of this network involves a clique of four nodes (molecules) with six overlapping reactions between them! To recognize such a circuit, where dynamics between a given set of molecules are best explained by multiple reactions, we must be able to decompose observed sensitivity profiles into additive combinations of smaller components, each of which corresponds to a basic reaction. The second problem is applicable in situations where reaction rates do not fall into the two basic types studied here. For instance, rate laws can be highly non-linear and involve more than one enzyme to catalyze a given reaction. Further, very fast rate constants can cause drastic changes in concentrations, too quick to be detectable by analyzing data.

Both these problems can be alleviated by numerical modeling of sensitivity profiles rather than the discrete approach of sensitivity tables as studied here. For instance, numerical optimization can be used to find fits to parameterized reaction laws and by repeatedly modeling the residual, we can detect multiple reactions spanning a given set of molecules. The last two rows of our 'All but 2' sensitivity table (Table 3) provide a limited capability in this regard and which we have used in the studies described below.

Finally, we mention that, in real applications, data collected from wet-lab experiments always contain some errors. We have to be aware that these errors are usually much larger than the numerical errors in the case studies described below. However, one advantage of our algorithm is its robustness. We do not require an accurate measurement of the sensitivity, just the signs of the sensitivities (relative to our threshold of 10^{-8} which can be tuned based on reliability of the measurements).

7. EXPERIMENTAL RESULTS

Our experimental results are focused on reconstructing key CRNs underlying important biological processes (see Ta-

Table 4: Summary of CRNs reconstructed and evaluation statistics.

Model	# species	# reactions	Recall	Precision	ODE/sensitivity solution time (10^{-3}s)	CRN mining time (10^{-3}s)
CDC-Cyclin2 interaction loop (Fig. 5)	6	6	0.83	0.83	42.3	0.27
Arkin's computational circuit (Fig. 6)	7	6	1	1	167	0.51
Prokaryotic gene expression model	9	8	0.875	0.875	97.6	0.56
Frog egg extracts (Fig. 4)	8	8	0.75	0.857	58	0.38
Generic yeast cell cycle model (Fig. 7)	16	21	0.857	0.88	637	2.31

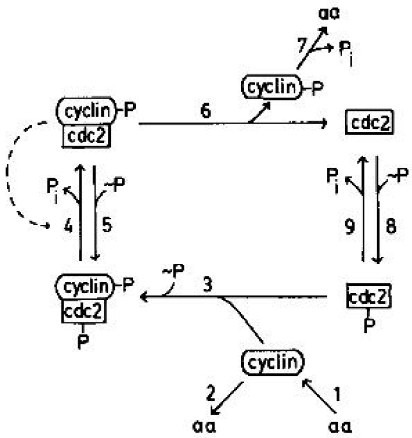


Figure 5: The CDC-Cyclin2 interaction loop forming the core of the budding yeast cell cycle. Courtesy John Tyson.

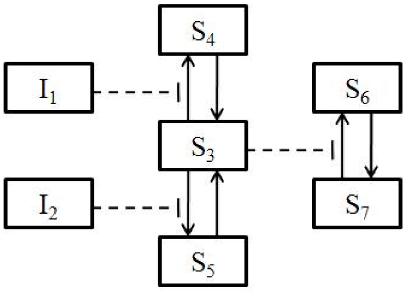


Figure 6: A CRN designed to serve as a computational element (i.e., as a logic gate).

ble 4). Here we depict the number of species and reactions for each system but hasten to add that the complexity of a CRN cannot be judged merely on these factors alone. For instance, the rather innocuous looking system from Fig. 1, referred to as the 'oregonator', forms the model for many reaction-diffusion systems and can exhibit very complex dynamics including sustained oscillations. It is hence the range of qualitative behaviors that can be exhibited by the system that constitutes its complexity.

For each CRN studied here, we formulated the corresponding ODE as described in Section 3, and generated data corresponding to each ODE using the CVODE software [6]. All rate law equations were modeled using either mass action kinetics or Michaelis Menten kinetics. For each pair of molecules, the buffering algorithm buffers all but these two molecules, and the sensitivity profiles between these

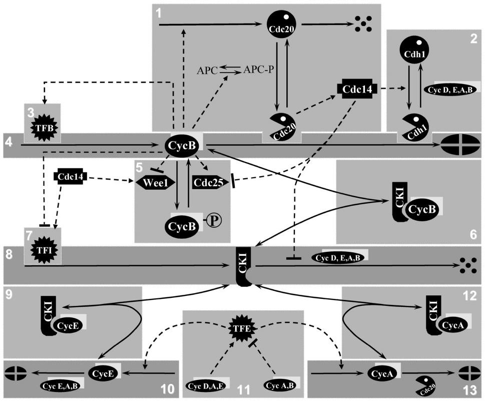


Figure 7: Generic CRN of the budding yeast cell cycle. Regulatory modules are given by the shaded rectangles. The different symbols denote different classes of proteins, e.g., the 'PacMan' denotes active forms of regulated proteins. Courtesy John Tyson.

molecules are computed. A tolerance of 10^{-8} was used to discretize the computed sensitivities. This information drives the reconstruction of topology and reaction characteristics. The results are evaluated using metrics of recall (number of correctly reconstructed reactions as a fraction of true reactions) and precision (number of correctly reconstructed reactions as a function of all reconstructed reactions). In assessing correctness, to allow partial matches, we evaluate reversible reactions in both directions (i.e., if the algorithm reconstructs the reaction in only one direction, we count it as one out of two reactions inferred correctly).

The CRNs considered here span a variety of model systems in biology. The CDC-Cyclin2 interaction loop (Fig. 5 [17]) is the core signaling pathway driving progression through the cell cycle. It is embedded inside the larger yeast cell cycle model described in Fig. 7 [7]. A less complex model drives cell cycle transitions in frog egg extracts, as described earlier in Fig. 4. Two other models considered here are a CRN underlying gene expression regulation in prokaryotes, which are primitive organisms such as bacteria that do not contain membrane-bound organelles (not shown due to space considerations) and a CRN meant to serve as a generic logic gate (Fig. 6).

As Table 4 reveals, our algorithm achieves consistently high values of recall and precision across these CRNs. The three reasons it fails to find correct reactions or infers spurious reactions are: the inherent inability to distinguish between certain types of reactions (as discussed earlier), rapid reaction rates that mistakenly cause the algorithm to infer lack of connectivity between some species, and the restriction to at most one reaction between a given pair or triple of molecules. Even with these caveats, it is clear that the algorithm can be used as a primitive to identify key circuits underlying a collection of molecules.

Table 4 also tabulates the time taken to reconstruct each CRN along with the time taken to solve the ODE as well as the associated buffering/sensitivity analysis experiments. Observe that the latter is a function of not just the size of the CRN but also the stiffness of the underlying ODE. (A stiff equation requires that the ODE integrator use an extremely small stepsize due to components varying at different time scales or because of underlying numerical instability.)

8. DISCUSSION

We have presented a novel application of data mining methodology to chemical reaction system identification with a marriage of numerical methods and graphical models. Our work is the first to address CRN mining using KDD concepts and methodology. The $O(n^2)$ buffering experiments required for our algorithm is not a severe constraint and special purpose combinatorial equipment can be utilized in larger systems. The supplementary website http://bioinformatics.cs.vt.edu/CRNMining provides sufficient details to reproduce the experiments described here.

Our future work focuses on three directions. First, we would like to employ an *Apriori* like approach to searching for groups of reactions in a given sensitivity matrix, so that if a given reaction can be ruled out from being present, so can all its supersets. However, this requires careful understanding of the areas where monotonicity constraints over the dynamics of CRNs apply. Second, we desire to connect our work better to theories of system identification, especially as a way to control the complexity of network reconstruction. Our work has focused exclusively on the time domain and more powerful analysis tools that work in the frequency domain can be brought to bear here. Finally, we wish to use our data mining algorithm as a aid to *network comprehension*, i.e., to summarize a complex CRN in terms of its information processing capabilities. For instance, groups of chemical reactions can be viewed as forming switches, amplifiers, or signal transducers [18]. By directly recognizing such circuit motifs, we can aid in reconstructing not just the structure of CRNs but their functional aspects as well.

9. REFERENCES

[1] A. Arkin, P. Shen, and J. Ross. A Test Case of Correlation Metric Construction of a Reaction Pathway from Measurements. *Science*, Vol. 277(5330):1275–1279, Aug 1997.

[2] U.S. Bhalla. Understanding Complex Signaling Networks through Models and Metaphors. *Progress in Biophysics and Molecular Biology*, Vol. 81(1):45–65, Jan 2003.

[3] F. Bromberg, D. Margaritis, and V. Honavar. Efficient Markov Network Structure Discovery using Independence Tests. In *Proceedings of the Sixth SIAM International Conference on Data Mining.* SIAM Press, 2006.

[4] P.N. Brown, A.C. Hindmarsh, and L.R. Petzold. Using Krylov Methods in the Solution of Large-Scale Differential-Algebraic Systems. *SIAM Journal of Scientific Computing*, Vol. 15:1467–1488, 1994.

[5] K.C. Chen, L. Calzone, F.R. Csikasz-Nagy, F.R. Cross, B. Novak, and J.J. Tyson. Integrative Analysis of Cell Cycle Control in Budding Yeast. *Molecular Biology of the Cell*, Vol. 15:3841–3862, Aug 2004.

[6] S. Cohen and A. Hindmarsh. CVODE, A Stiff/Nonstiff ODE Solver in C. *Computers in Physics*, Vol. 10(2):138–143, Mar-Apr 1996.

[7] A. Csikasz-Nagy, D. Battogtokh, K.C. Chen, B. Novak, and J.J. Tyson. Analysis of a Generic Model of Eukaryotic Cell Cycle Regulation. *Biophys. J.*, Vol. 90:4361–4379, 2006.

[8] A.V. Karnaukhov and E.V. Karnaukhova. System Identification in Biophysics: A New Method based on Minimizing Square Residuals. *Biofizika*, Vol. 49(1):88–97, 2004.

[9] A.V. Karnaukhov, E.V. Karnaukhova, and J.R. Williamson. Numerical Matrices Method for Nonlinear System Identification and Description of Dynamics of Biochemical Reaction Networks. *Biophysical Journal*, Vol. 92:3459–3473, 2007.

[10] S. Klamt and E.D. Gilles. Minimal Cut Sets in Biochemical Reaction Networks. *Bioinformatics*, Vol. 20(2):226–234, 2004.

[11] E. Klipp, R. Herwig, A. Kowald, C. Wierling, and H. Lehrach. *Systems Biology in Practice.* Wiley-VCH, May 2005.

[12] S.L. Lauritzen. *Graphical Models.* Oxford University Press, 1996.

[13] W. Marwan, A. Wagler, and R. Weismantel. A Mathematical Approach to Solve the Network Reconstruction Problem. *Mathematical Methods in Operations Research*, Vol. 67:117–132, 2008.

[14] M.R. Maurya, S.J. Bornheimer, V. Venkatasubramanian, and S. Subramaniam. Reduced-order Modelling of Biochemical Networks: Application to the GTpase-Cycle Signalling Module. *IEE Systems Biology*, Vol. 152(4):229–242, Dec 2005.

[15] J. Ross, I. Schreiber, and M.O. Vlad. *Determination of Complex Reaction Mechanisms: Analysis of Chemical, Biological, and Genetic Networks.* Oxford University Press, Nov 2005.

[16] W. Sha, J. Moore, K. Chen, A.D. Lassaletta, C.-S. Yi, J.J. Tyson, and J. Sible. Hysteresis drives Cell-cycle Transitions in *Xenopus laevis* Egg Extracts. *PNAS*, Vol. 100(3):975–980, Feb 2003.

[17] J.J. Tyson. Modeling the Cell Division Cycle: cdc2 and Cyclin Interactions. *PNAS*, Vol. 88(16):7328–7332, Aug 1991.

[18] J.J. Tyson, K.C. Chen, and B. Novak. Sniffers, Buzzers, Toggles and Blinkers: Dynamics of Regulatory and Signaling Pathways in the Cell. *Current Opinion in Cell Biology*, Vol. 15(2):221–231, Apr 2003.

[19] C.H. Wiggins and I. Nemenman. Process Pathway Inference via Time Series Analysis. *Experimental Mechanics*, Vol. 43(3):361–370, Sep 2003.