International Scholarly Research Network
ISRN Biomathematics
Volume 2012, Article ID 613174, 7 pages
doi:10.5402/2012/613174

Research Article

Inferring Biologically Relevant Models: Nested Canalyzing Functions

Franziska Hinkelmann[1] and Abdul Salam Jarrah[2]

[1] *Mathematical Biosciences Institute, Ohio State University, Columbus, OH 43210, USA*
[2] *Department of Mathematics and Statistics, American University of Sharjah, P.O. Box 26666, Sharjah, United Arab Emirates*

Correspondence should be addressed to Abdul Salam Jarrah, ajarrah@aus.edu

Received 9 March 2012; Accepted 9 April 2012

Academic Editors: J. Chow and J. Fellman

Inferring dynamic biochemical networks is one of the main challenges in systems biology. Given experimental data, the objective is to identify the rules of interaction among the different entities of the network. However, the number of possible models fitting the available data is huge, and identifying a biologically relevant model is of great interest. Nested canalyzing functions, where variables in a given order dominate the function, have recently been proposed as a framework for modeling gene regulatory networks. Previously, we described this class of functions as an algebraic toric variety. In this paper, we present an algorithm that identifies all nested canalyzing models that fit the given data. We demonstrate our methods using a well-known Boolean model of the cell cycle in budding yeast.

1. Introduction

Inferring dynamic biochemical networks is one of the main challenges in systems biology. Many mathematical and statistical methods, within different frameworks, have been developed to address this problem, see [1] for a review of some of these methods. Starting from experimental data and known biological properties only, the idea is to infer a "most likely" model that could be used to generate the experimental data. Here the model could have two parts. The first one is the static network which is a directed graph showing the influence relationships among the components of the network, where an edge from node y to node x implies that changes in the concentration of y could change the concentration of x. The other part of the model is the dynamics of the network, which describes how exactly the concentration of x is affected by that of y. Due to the fact that biological networks are not well understood and the available data about the network is usually limited, many models end up fitting the available information, and the criteria for choosing a particular model are usually not biologically motivated but rather a consequence of the modeling framework.

A framework that has long been used for modeling gene regulatory networks is *time-discrete, finite-space dynamical systems*. This includes Boolean networks [2], Logical models [3], Petri nets [4], and algebraic models [5]. The latter is a straightforward generalization of Boolean networks to multistate systems. Furthermore, in [6], it was shown that logical models as well as Petri nets could be viewed and analyzed as algebraic models. The inference methods we develop here are within the algebraic models framework. To be self-contained, we briefly describe this framework and state some of the known results that we need in this paper, see [5–8] for more details. Throughout this paper, we will be talking about gene regulatory networks; however, the methods apply to biochemical networks in general.

Suppose that the gene regulatory network that we want to infer has n genes and that we have a set D of r state transition pairs $(\mathbf{s}_j, \mathbf{t}_j)$, $j = 1, \ldots, r$. The input $\mathbf{s}_j$ and the output $\mathbf{t}_j$ are n-tuples of 0 and 1 encoding the state of genes $x_1, \ldots, x_n$. Real-time data points are not Boolean but could be discretized (and in particular, could be made Boolean) using different methods [9]. The goal is to find a model:

$$f = (f_1, f_2, \ldots, f_n) : \mathbb{F}_2^n \longrightarrow \mathbb{F}_2^n, \tag{1}$$

such that for $j = 1, \ldots, r$:

$$f(\mathbf{s}_j) = \left(f_1(\mathbf{s}_j), \ldots, f_n(\mathbf{s}_j)\right) = \mathbf{t}_j. \quad (2)$$

Notice that, since any function over a finite field is a polynomial, each f_i is a polynomial. An algorithm that finds all models f is presented in [5]. This is done by identifying, for each gene i, the set of all possible functions for f_i. This set can be represented as the coset $f + I$, where f is a particular such function, and $I \subset \mathbb{F}_2[x_1, \ldots, x_n]$ is the ideal of all Boolean polynomials that vanish on the input data set, that is, $I = \mathbb{I}(\{\mathbf{s}_1, \ldots, \mathbf{s}_r\})$. The algorithm in [5] then proceeds to find a particular model from the model space $f + I$. The chosen model, which is the normal form of f in the ideal I, depends on the term ordering used in the Gröbner bases computation. So different ordering of the variables (genes) might lead to the selection of different models. This presents a problem as the term ordering, which is a needed for computational reasons, clearly influence the model selection process.

Several modifications have since been presented to address this problem. For example, in [10], using the Gröbner fan of the ideal I, the authors developed a method that produces a probabilistic model using all possible normal forms. Other improvements on this algorithm can be found in [11, 12].

Another approach toward improving the model selection process is by restricting the model space $f + I$ by requiring not only that the chosen model fits the data but also satisfing some other conditions, such as its network being sparse or scale-free, the polynomials f_i being monomials, the dynamics of the model having some desirable properties such as fixed points are the only limit cycles (i.e., starting from any initialization, the model always reaches a steady state), or that the model is robust and stable which could roughly mean that the number of attractors in the phase space is small. In a nutshell, some but not all functions in the model space $f + I$ are biologically relevant, and hence restricting the space to only relevant models will improve the model selection process.

By desiring a particular property, several classes of functions have been proposed as biologically relevant functions such as biologically meaningful rules [13], certain postclasses of Boolean functions have been studied in [14], and chain functions in [15], to name few. Another class of Boolean functions, which was introduced by Kauffman et al. [16], is called (nested) canalyzing functions (NCF), where an input to a single variable exclusively determines the value of the function regardless of the values of all other variables. This is a natural characterization of "canalisation" which was introduced by geneticist Waddington [17] to represent the ability of a genotype to produce the same phenotype regardless of environmental variability. Indeed, known biological functions have been shown to be canalyzing [18, 19], and Boolean nested canalyzing networks to be robust and stable [16, 19, 20].

For the purpose of restricting the model space $f + I$ of all Boolean polynomial models to NCFs only, we previously studied nested canalyzing functions, gave necessary and sufficient conditions on the coefficients of a boolean polynomial function to be nested canalyzing, and showed that NCFs are nothing but unate cascade functions [7]. Furthermore, in [8], the class of all nested canalyzing functions is parameterized as the rational points of an affine algebraic variety over the algebraic closure of $\mathbb{F}_2$. This variety was shown to be toric, that is, defined by a collection of binomial polynomial equations. In this paper, we present an algorithm that restricts the model space to only nested canalyzing functions by identifying all NCFs from the model space $f + I$ that fit the given data set.

In the next section, we briefly recall some definitions and results from [7, 8]. Our algorithm is presented in Section 3, and its implementation in Singular is discussed in Section 4. Before we conclude this paper, we demonstrate the algorithm in Section 5, where we identify all nested canalyzing models for the cell cycle in budding yeast using time course data from the Boolean model in [21].

2. Nested Canalyzing Functions: Background

We recall some of the definitions and the results from [7, 8] that we need to make this presentation self-contained. Throughout this paper, when we refer to a function on n variables, we mean that h depends on all n variables, that is, for $i = 1, \ldots, n$, there exists $(a_1, \ldots, a_n) \in \mathbb{F}_2^n$ such that $h(a_1, \ldots, a_{i-1}, a_i, a_{i+1}, \ldots, a_n) \neq h(a_1, \ldots, a_{i-1}, 1 + a_i, a_{i+1}, \ldots, a_n)$.

Definition 1. Let h be a Boolean function on n variables, that is, $h : \mathbb{F}_2^n \to \mathbb{F}_2$:

(i) the function h is a *nested canalyzing function* (NCF) with respect to a permutation σ on the n variables, canalyzing input value a_i and canalyzed output value b_i, for $i = 1, \ldots, n$, if it can be represented in the form:

$$h(x_1, x_2, \ldots, x_n) = \begin{cases} b_1 & \text{if } x_{\sigma(1)} = a_1, \\ b_2 & \text{if } x_{\sigma(1)} \neq a_1, x_{\sigma(2)} = a_2, \\ b_3 & \text{if } x_{\sigma(1)} \neq a_1, x_{\sigma(2)} \neq a_2, \\ & \quad x_{\sigma(3)} = a_3, \\ \vdots & \vdots \\ b_n & \text{if } x_{\sigma(1)} \neq a_1, \ldots, x_{\sigma(n-1)} \neq a_{n-1}, \\ & \quad x_{\sigma(n)} = a_n, \\ \overline{b_n} & \text{if } x_{\sigma(1)} \neq a_1, \ldots, x_{\sigma(n)} \neq a_n; \end{cases} \quad (3)$$

(ii) the function h is *nested canalyzing* if h is nested canalyzing with respect to some permutation σ, canalyzing input values $a_1, \ldots, a_n$ and canalyzed output values $b_1, \ldots, b_n$, respectively.

Remark 2. The definition above has been generalized to multistate functions in [22], where it is also shown that the dynamics of these functions are similar to their Boolean counterparts. In [23], the authors introduce what they called

kinetic models with unate structure, which are continuous models having the canalization property, and they presented an algorithm for identifying such models.

Using the polynomial form of any Boolean function, the ring of Boolean functions is isomorphic to the quotient ring $R = \mathbb{F}_2[x_1, \ldots, x_n]/J$, where $J = \langle x_i^2 - x_i : 1 \le i \le n \rangle$. Indexing monomials by the subsets of $[n] := \{1, \ldots, n\}$ corresponding to the variables appearing in the monomial, the elements of R can be written as

$$R = \left\{ \sum_{S \subseteq [n]} c_S \prod_{i \in S} x_i : \ c_S \in \mathbb{F}_2 \right\}. \tag{4}$$

As a vector space over $\mathbb{F}_2$, R is isomorphic to $\mathbb{F}_2^{2^n}$ via the correspondence:

$$R \ni \sum_{S \subseteq [n]} c_S \prod_{i \in S} x_i \longleftrightarrow (c_\emptyset, \ldots, c_{[n]}) \in \mathbb{F}_2^{2^n}. \tag{5}$$

The main result in [7] is the identification of the set of nested canalyzing functions in R with a subset V^{ncf} of $\mathbb{F}_2^{2^n}$ by imposing relations on the coordinates of its elements.

Definition 3. Let σ be a permutation of the elements of the set $[n]$. We define a new order relation $<_\sigma$ on the elements of $[n]$ as follows: $\sigma(i) <_\sigma \sigma(j)$ if and only if $i < j$. Let r_S^σ be the maximum element of a nonempty subset S of $[n]$ with respect to the order relation $<_\sigma$. For any nonempty subset S of $[n]$, the *completion of S with respect to the permutation σ*, denoted by $[r_S^\sigma]$, is the set $[r_S^\sigma] = \{\sigma(1), \sigma(2), \ldots, \sigma(r_S^\sigma)\}$.

Note that, if σ is the identity permutation, then the completion is $[r_S] := \{1, 2, \ldots, r_S\}$, where r_S is the largest element of S.

Theorem 4. *Let $h \in R$, and let σ be a permutation of the set $[n]$. The polynomial h is nested canalyzing with respect to σ, input value a_i and corresponding output value b_i, for $i = 1, \ldots, n$, if and only if $c_{[n]} = 1$ and, for any proper subset $S \subseteq [n]$:*

$$c_S = c_{[r_S^\sigma]} \prod_{\sigma(i) \in [r_S^\sigma] \setminus S} c_{[n] \setminus \{\sigma(i)\}}. \tag{6}$$

Corollary 5. *The set of points in $\mathbb{F}_2^{2^n}$ corresponding to the set of all nested canalyzing functions with respect to a permutation σ on $[n]$, denoted by V_σ^{ncf}, is defined by*

$$V_\sigma^{\mathrm{ncf}} = \left\{ (c_\emptyset, \ldots, c_{[n]}) \in \mathbb{F}_2^{2^n} : c_{[n]} = 1, \ c_S = c_{[r_S^\sigma]} \times \prod_{\sigma(i) \in [r_S^\sigma] \setminus S} c_{[n] \setminus \{\sigma(i)\}}, \ \text{for } S \subseteq [n] \right\}. \tag{7}$$

It was shown in [8] that V_σ^{ncf} is an algebraic variety, and its ideal $\mathbb{I}(V_\sigma^{\mathrm{ncf}})$ is a binomial prime ideal in the polynomial ring $\overline{\mathbb{F}}_2[\{c_S : S \subseteq [n]\}]$, where $\overline{\mathbb{F}}_2$ is the algebraic closure of $\mathbb{F}_2$. Namely,

$$\begin{aligned} I_\sigma &= \mathbb{I}\left(V_\sigma^{\mathrm{ncf}}\right) \\ &= \left\langle c_{[n]} - 1, \ c_S - c_{[r_S^\sigma]} \prod_{\sigma(i) \in [r_S^\sigma] \setminus S} c_{[n] \setminus \{\sigma(i)\}} : S \subset [n] \right\rangle. \end{aligned} \tag{8}$$

Furthermore, the variety of all nested canalyzing functions is

$$V^{\mathrm{ncf}} = \bigcup_\sigma V_\sigma^{\mathrm{ncf}}, \tag{9}$$

and its ideal is

$$\mathbb{I}\left(V^{\mathrm{ncf}}\right) = \bigcap_\sigma I_\sigma. \tag{10}$$

In the next section, we identify the set $f + I$ with the rational points in an algebraic affine variety. This will allow us to identify all nested canalyzing functions in the model space $f + I$.

3. Nested Canalyzing Models

Recall that we are given the data set $D = \{(\mathbf{s}_1, \mathbf{t}_1), \ldots, (\mathbf{s}_r, \mathbf{t}_r)\} \subset \mathbb{F}_2^n \times \mathbb{F}_2^n$. The model space could be presented by the set $f + I$, where $f = (f_1, \ldots, f_n)$ and, for $i = 1, \ldots, n$,

$$f_i(x_1, \ldots, x_n) = \sum_{j=1}^{r} t_{j,i} \prod_{e=1}^{n} \left(1 - \left(x_e - s_{j,e}\right)\right). \tag{11}$$

In particular, f_i is a polynomial that interpolates the data for gene i and I is the *ideal of points* of $\{\mathbf{s}_1, \ldots, \mathbf{s}_r\}$. Furthermore, the ideal I is a principal ideal in the ring R/J:

$$\begin{aligned} I &= \mathbb{I}(\{\mathbf{s}_1, \ldots, \mathbf{s}_r\}) \\ &= \bigcap_{j=1}^{r} \mathbb{I}\left(\left\{\mathbf{s}_j\right\}\right) \\ &= \bigcap_{j=1}^{r} \left\langle x_1 - s_{j,1}, \ldots, x_n - s_{j,n} \right\rangle \\ &= \bigcap_{j=1}^{r} \left\langle 1 - \prod_{e=1}^{n} \left(1 - \left(x_e - s_{j,e}\right)\right) \right\rangle \\ &= \left\langle \prod_{j=1}^{r} \left(1 - \prod_{\mathrm{e}=1}^{n} \left(1 - \left(x_e - s_{j,e}\right)\right) \right) \right\rangle. \end{aligned} \tag{12}$$

Now a polynomial $h \in f_i + I$ if and only if $h = f_i + g(x_1, \ldots, x_n) \prod_{j=1}^{r} (1 - \prod_{e=1}^{n} (1 - (x_e - \mathbf{s}_{j,e})))$, for some polynomial g, say $g = \sum_{H \subseteq [n]} b_H \prod_{i \in H} x_i$. By expanding the right-hand side and collecting terms, we get that $h = \sum_{S \subseteq [n]} W_S(b_H, \mathbf{s}_j, \mathbf{t}_j) \prod_{l \in S} x_l$, where for $S \subseteq [n]$, the coefficient $W_S(b_H, \mathbf{s}_j, \mathbf{t}_j)$ is determined by $b_H, \mathbf{s}_j, \mathbf{t}_j$ for all $H \subseteq [n]$ and $j = 1, \ldots, r$.

The proof of the following theorem follows directly from Theorem 2.4.2 in [24].

Theorem 6. *Consider the ring homomorphism:*

$$\Phi : \overline{\mathbb{F}}_2[\{c_S : S \subseteq [n]\}] \longrightarrow \overline{\mathbb{F}}_2[\{b_H : H \subseteq [n]\}] \tag{13}$$

given by, for $S \subseteq [n]$,

$$c_S \longmapsto W_S\left(b_H, \mathbf{s}_j, \mathbf{t}_j\right). \tag{14}$$

Then $\ker(\Phi)$ *is the ideal of all polynomials that fit the data set D. In particular, the rational points in the variety* $\mathbb{V}(\ker(\Phi))$ *is the set of all models that fit the data set D, namely* $f + I$.

Since the ideal of all NCFs is $\mathbb{I}(V^{\mathrm{ncf}})$, the following corollary is straightforward.

Corollary 7. *The ideal of all nested canalyzing functions that fit the data set D is* $\mathbb{I}(V^{\mathrm{ncf}}) + \ker(\Phi)$.

Remark 8. It is clear that the model space of Boolean functions is huge, since the number of monomials grows exponentially in the number of variables. For example, if a function has 5 inputs, there are $2^5 = 32$ different monomials in 5 variables, and hence $2^{32} = 4,294,967,296$ different Boolean functions. This clearly shows that a search for NCFs inside the model space is computationally not feasible, which justifies the need for algorithms like the one above.

4. Algorithm

In this section, we present an algorithm for identifying all nested canalyzing models from the model space of a given data set.

Input. A wiring diagram, that is, a square matrix of dimension n, describing the influence relationships among the n genes in the network. For each variable x_i, a table consisting of the rows $(\mathbf{s}_{j,i_1}, \ldots, \mathbf{s}_{j,i_s}, \mathbf{t}_{j,i})$, $j = \{1, \ldots, r\}$, where $i_1, \ldots, i_s$ are the indices of the genes that affect x_i, as specified in the wiring diagram.

Output. For each variable, the complete list of all nested canalyzing functions interpolating the given data set on the given wiring diagram. A function is in the output if it is nested canalyzing in at least one variable order. If needed, the code can easily be modified to find only nested canalyzing functions of a particular variable order.

Algorithm 9. It is a well-known fact, that a Gröbner basis for the kernel of Φ is a basis for $\langle\{c_S - W_S : S \subseteq [n]\}\rangle$ intersected with the ring $\overline{F}_2[\{c_S : S \subseteq [n]\}]$ ([24], Theorem 2.4.2) Using a similar notation as above, the algorithm is outlined as follows:

use ring $\mathbb{F}_2[\{x_1, \ldots, x_n, b_S, c_S : S \subseteq [n]\}]$;
define $\mathbb{I}(V^{\mathrm{ncf}})$ as ideal in $\mathbb{F}_2[\{c_S : S \subseteq [n]\}]$;
define $h = \sum_{H \subseteq [n]} b_H \prod_{i \in H} x_i$;
define $q = \sum_{H \subseteq [n]} c_H \prod_{i \in H} x_i$;
compute the polynomial p that generates I as in (12);
for each variable x_i do

(1) compute f_i as in (11);
(2) let $g = f_i + h * p$; its coefficients are the same as W_S above;
(3) compute a Gröbner basis G for the ideal generated by the coefficients of $g - q$ using any elimination order to eliminate all b_S from G;
(4) concatenate generators of G and $\mathbb{I}(V^{\mathrm{ncf}})$;
(5) compute the primary decomposition of $G + \mathbb{I}(V^{\mathrm{ncf}})$ to obtain necessary and sufficient conditions on the coefficients of all NCFs fitting the data set D;

End.

An implementation of this algorithm is available as a Singular library [25, 26].

5. Application: Inferring the Cell Cycle Network in Budding Yeast

The interactions among proteins constitute complex molecular networks that regulate cell behavior such as the decision to undergo cell division. We use data points generated by a previously published model of the cell-cycle regulatory network of the budding yeast *Saccharomyces cerevisiae* to demonstrate the algorithm described in this paper [21]. The model is a Boolean model, where nodes represent proteins and edges describe protein interactions that are either activating or inhibiting. Proteins are updated according to a *threshold* rule, that is, a protein is activated, respectively deactivated, if the weighted sum of the activating input proteins is greater than, respectively less than, the weighted sum of the inhibiting input proteins. The model contains the key regulators of the cell cycle process and the known interactions among these regulators. This model captures the known features of the global dynamics of the cell cycle, it is robust and stable, and the trajectory of the known cell-cycle sequence is a stable and attracting trajectory as it has 1764 states out of the total number of 2048 states. The remaining states are distributed into 6 small trajectories.

In this section, we use the time course corresponding to the biological cell-cycle sequence, see Table 1, to infer nested canalyzing models of the cell cycle. That is, assuming the same wiring diagram as the threshold model in [21], we use our algorithm to identify, for each gene in the network, all nested canalyzing functions that fit the cell cycle sequence.

We start by describing the model. There are twelve nodes that represent eleven proteins and a start signal. The proteins are members of the following three classes: cyclins (Cln1,2, Cln3, Clb1,2, and Clb5,6), inhibitors, degraders, and competitors of cyclin complexes (Sic1, Cdh1, Cdc20, and Cdc14), and transcription factors (SBF, MBF, Swi5, and Mcm1/SFF). This simplified network (Figure 1) is almost identical to the network in [21], where the only difference is that we do not force self-degradation, as it was added to some

Table 1: The temporal evolution of the Boolean cell-cycle model in [21]; corresponding to the biological cell-cycle sequence.

Time	Cln3	MBF	SBF	Cln1,2	Cdh1	Swi5	Cdc14,20	Clb5,6	Sic1	Clb1,2	Mcm1/SFF
1	1	0	0	0	1	0	0	0	1	0	0
2	0	1	1	0	1	0	0	0	1	0	0
3	0	1	1	1	1	0	0	0	1	0	0
4	0	1	1	1	0	0	0	0	0	0	0
5	0	1	1	1	0	0	0	1	0	0	0
6	0	1	1	1	0	0	0	1	0	1	1
7	0	0	0	1	0	0	1	1	0	1	1
8	0	0	0	0	0	1	1	0	0	1	1
9	0	0	0	0	0	1	1	0	1	1	1
10	0	0	0	0	0	1	1	0	1	0	1
11	0	0	0	0	1	1	1	0	1	0	0
12	0	0	0	0	1	1	0	0	1	0	0
13	0	0	0	0	1	0	0	0	1	0	0

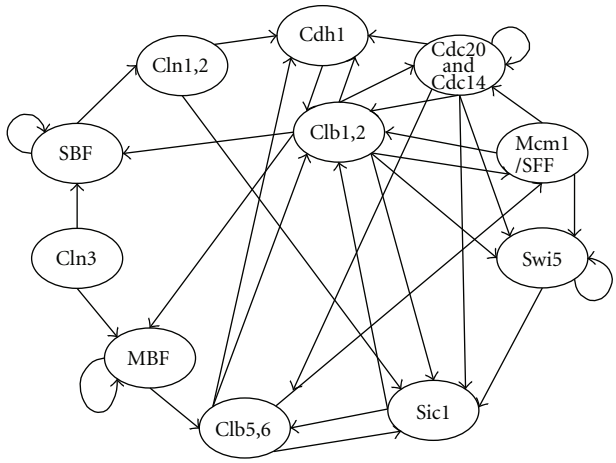

Figure 1: The simplified cell cycle network in budding yeast, which is based on the model in [21].

nodes in the network because they did not have inhibitors, but without biological justification [21]. Furthermore, we do not impose activation or inhibition in the network. As we do not use threshold functions but more general boolean functions, a variable can both increase and decrease the concentration of another substrate, depending on the concentrations of other proteins.

Li et al. [21] use their model to generate a time course of temporal evolution of the cell-cycle network, shown in Table 1. This time course is in agreement with the behavior of the cell-division process cycling through the four distinct phases G_1, S (synthesis), G_2, and M (mitosis).

We used this time course along with the network as the only input to the algorithm to obtain all nested canalyzing functions that interpolate the time course. In the fifth column of Table 2, we list the number of NCFs for each protein. By requiring the Boolean function to be nested canalyzing, we have significantly reduced the number of possible functions for each protein as it is evident when comparing the numbers in the third and fifth columns of Table 2. However, even after this reduction, there are 330, 559, 488 possible nested canalyzing models that fit the time course in Table 1. To reduce this number more, one needs to use additional time courses or request that the models incorporate additional biological information about yeast cell cycle.

5.1. Dynamics. To analyze the dynamics of the resulting nested canalyzing models, we randomly sampled 2000 models and analyzed them. The average number of basins of attraction (components) per network is 3.09, and the average size of the component containing the given trajectory is 1889. In only 6 models, the trajectory in Table 1 is not in the largest component; however, the average size of the component containing the trajectory is 833.5.

These results clearly show that nested canalyzing models for the cell cycle network are in agreement with the original threshold model of Li et al., and since such models are known to be robust and stable, any of these models could be used as a model for the cell cycle in budding yeast. Furthermore, especially when there is no evidence for choosing a particular type of functions, a nested canalyzing function has an advantage over other possible choices.

5.2. Comparison with Random Networks. To understand the effect of the network itself on its dynamics, we sampled 2000 models on the same network, where the local function of each gene in each one of these models is chosen randomly from all possible functions in the model space. We found that the given cell cycle trajectory has oftentimes much smaller basin of attraction, and hence random functions on the cell cycle network could not in general produce the desired dynamics. A comparison of the statistics from the sampled networks is shown in Figures 2 and 3.

6. Conclusion

In this paper, we have presented an algorithm for identifying all Boolean nested canalyzing models that fit a given time course or other input-output data sets. Our algorithm uses

Table 2: For each protein i, we list the number of inputs, the number of possible Boolean functions (the cardinality of $f_i + I$), the number of nested canalyzing functions with the given number of inputs, and finally the number of nested canalyzing functions in the model space $f_i + I$.

Protein (i)	Inputs	$f_i + I$	NCFs	NCFs in $f_i + I$
Cln3	1	1	2	1
MBF	3	8	64	2
SBF	3	8	64	2
Cln1,2	1	1	2	1
Cdh1	4	2048	736	12
Swi5	4	2048	736	14
Cdc20 and Cdc14	3	8	64	4
Clb5,6	3	8	64	3
Sic1	5	2^{24}	10,634	336
Clb1,2	5	2^{24}	10,634	61
Mcm1/SFF	3	8	64	2

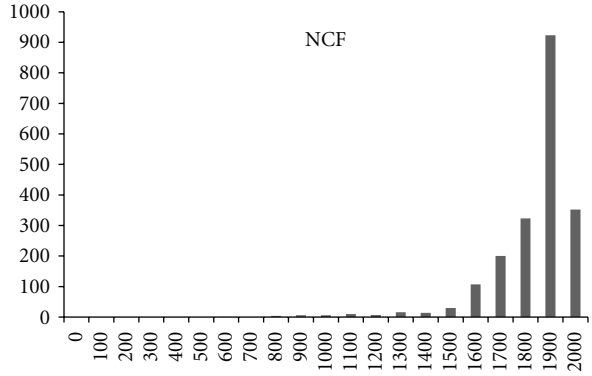

Figure 2: Nested canalyzing functions with the wiring diagram in Figure 1 interpolating the time course in Table 1. x-axis: size of basin of attraction for given trajectory; y-axis: number of networks observed, out of 2000.

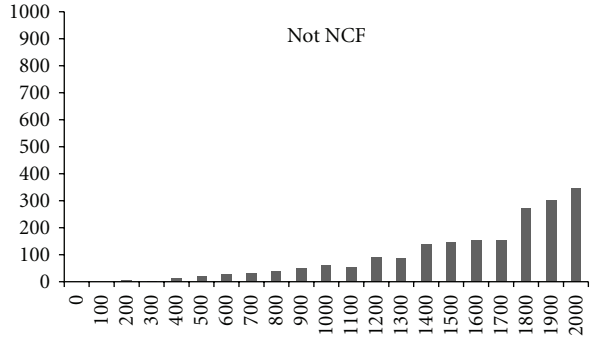

Figure 3: Not nested canalyzing functions with the wiring diagram in Figure 1 interpolating the time course in Table 1. x-axis: size of basin of attraction for given trajectory; y-axis: number of networks observed, out of 2000.

methods from computational algebra to present the model space as an algebraic variety. The intersection of this variety with the variety of all NCFs, which was parameterized in [8], gives us the set of all NCFs that fit the data. We demonstrate our algorithm by finding all nested canalyzing models of the cell cycle network form Li et al. [21]. We then showed that the dynamics of almost any of these models is strikingly similar to that of the original threshold model. Unless the chosen model is required to meet other conditions, and in that case the model space will be reduced further, any one of the models that our algorithm found is an acceptable model of the cell cycle process in the budding yeast.

One limitation of the current algorithm, which we left for future work, is that it does not distinguish between activation and inhibition in the network as we do not have a systematic method of knowing when a given variable in a (nested canalyzing) polynomial is an activator or inhibitor.

As our algorithm relies heavily on different Gröbner-based computations, the current implementation in Singular allows a given gene to have at most 5 regulators. This is due to the fact that the number of monomials then is 32 which is already a burden especially when the primary decomposition of an ideal is what we are after. We are working on a better implementation so that we can infer larger and denser networks.

Acknowledgments

F. Hinkelmann was supported by a grant from the U.S. Army Research Office. The authors thank the referees for their comments and suggestions.

References

[1] C. Sima, J. Hua, and S. Jung, "Inference of gene regulatory networks using time-series data: a survey," *Current Genomics*, vol. 10, no. 6, pp. 416–429, 2009.

[2] S. A. Kauffman, "Metabolic stability and epigenesis in randomly constructed genetic nets," *Journal of Theoretical Biology*, vol. 22, no. 3, pp. 437–467, 1969.

[3] R. Thomas, "Boolean formalization of genetic control circuits," *Journal of Theoretical Biology*, vol. 42, no. 3, pp. 563–585, 1973.

[4] L. J. Steggles, R. Banks, O. Shaw, and A. Wipat, "Qualitatively modelling and analysing genetic regulatory networks: A Petri net approach," *Bioinformatics*, vol. 23, no. 3, pp. 336–343, 2007.

[5] R. Laubenbacher and B. Stigler, "A computational algebra approach to the reverse engineering of gene regulatory networks," *Journal of Theoretical Biology*, vol. 229, no. 4, pp. 523–537, 2004.

[6] A. Veliz-Cuba, A. S. Jarrah, and R. Laubenbacher, "Polynomial algebra of discrete models in systems biology," *Bioinformatics*, vol. 26, no. 13, Article ID btq240, pp. 1637–1643, 2010.

[7] A. S. Jarrah, B. Raposa, and R. Laubenbacher, "Nested canalyzing, unate cascade, and polynomial functions," *Physica D*, vol. 233, no. 2, pp. 167–174, 2007.

[8] A. Jarrah and R. Laubenbacher, "Discrete models of biochemical networks: the toric variety of nested canalyzing functions," in *Algebraic Biology*, H. Anai, K. Horimoto, and T. Kutsia, Eds., vol. 4545 of *Lecture Notes in Computer Science*, pp. 15–22, Springer, 2007.

[9] E. S. Dimitrova, M. P. V. Licona, J. McGee, and R. Laubenbacher, "Discretization of time series data," *Journal of Computational Biology*, vol. 17, no. 6, pp. 853–868, 2010.

[10] E. S. Dimitrova, A. S. Jarrah, R. Laubenbacher, and B. Stigler, "A Gröbner fan method for biochemical network modeling," in *Proceedings of the International Symposium on Symbolic and Algebraic Computation (ISSAC '07)*, pp. 122–126, Ontario, Canada, August 2007.

[11] A. S. Jarrah, R. Laubenbacher, B. Stigler, and M. Stillman, "Reverse-engineering of polynomial dynamical systems," *Advances in Applied Mathematics*, vol. 39, no. 4, pp. 477–489, 2007.

[12] B. Stigler, A. Jarrah, M. Stillman, and R. Laubenbacher, "Reverse engineering of dynamic networks," *Annals of the New York Academy of Sciences*, vol. 1115, pp. 168–177, 2007.

[13] L. Raeymaekers, "Dynamics of Boolean networks controlled by biologically meaningful functions," *Journal of Theoretical Biology*, vol. 218, no. 3, pp. 331–341, 2002.

[14] I. Shmulevich, H. Lähdesmäki, E. R. Dougherty, J. Astola, and W. Zhang, "The role of certain post classes in Boolean network models of genetic networks," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 100, no. 19, pp. 10734–10739, 2003.

[15] I. Gat-Viks and R. Shamir, "Chain functions and scoring functions in genetic networks," *Bioinformatics*, vol. 19, no. 1, pp. 108–117, 2003.

[16] S. Kauffman, C. Peterson, B. Samuelsson, and C. Troein, "Genetic networks with canalyzing Boolean rules are always stable," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 101, no. 49, pp. 17102–17107, 2004.

[17] C. H. Waddington, "Canalization of development and the inheritance of acquired characters," *Nature*, vol. 150, no. 3811, pp. 563–565, 1942.

[18] S. Harris, B. Sawhill, A. Wuensche, and S. Kauffman, "A model of transcriptional regulatory networks based on biases in the observed regulation rules," *Complex*, vol. 7, no. 4, pp. 23–40, 2002.

[19] S. Nikolajewa, M. Friedel, and T. Wilhelm, "Boolean networks with biologically relevant rules show ordered behavior," *BioSystems*, vol. 90, no. 1, pp. 40–47, 2007.

[20] S. Kauffman, C. Peterson, B. Samuelssont, and C. Troein, "Random Boolean network models and the yeast transcriptional network," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 100, no. 25, pp. 14796–14799, 2003.

[21] F. Li, T. Long, Y. Lu, Q. Ouyang, and C. Tang, "The yeast cell-cycle network is robustly designed," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 101, no. 14, pp. 4781–4786, 2004.

[22] D. Murrugarra and R. Laubenbacher, "The number of multistate nested canalyzing functions," *Physica*, vol. 241, no. 10, pp. 929–938, 2012.

[23] R. Porreca, E. Cinquemani, J. Lygeros, and G. Ferrari-Trecate, "Identification of genetic network dynamics with unate structure," *Bioinformatics*, vol. 26, no. 9, Article ID btq120, pp. 1239–1245, 2010.

[24] W. Adams and P. Loustaunau, *An Introduction to Gröbner Bases*, vol. 3 of *Graduate Studies in Mathematics*, American Mathematical Society, Providence, RI, USA, 1994.

[25] F. Hinkelmann, "Singular implementation of NCF inferring algorithm," 2010, http://www.math.vt.edu/people/fhinkel/ncf.lib.

[26] W. Decker, G.-M. Greuel, G. Pfister, and H. Schönemann, "Singular 3-1-1—a computer algebra system for polynomial computations," 2010, http://www.singular.unikl.de.