JMLR: Workshop and Conference Proceedings 6: 215-224 NIPS 2008 workshop on causality

SIGNET: Boolean Rule Determination for Abscisic Acid Signaling

Jerry Jenkins JWJ@CFDRC.COM
Systems Biology and Bioinformatics Group
CFD Research Corporation
601 Genome Way
Huntsville, AL 35806, USA

Editor: Isabelle Guyon, Dominik Janzing, and Bernhard Schölkopf

Abstract

This paper describes the SIGNET dataset generated for the Causality Challenge. Cellular signaling pathways are most elusive types of networks to access experimentally due to the lack of methods for determining the state of a signaling network in an intact living cell. Boolean network models are currently being used for the modeling of signaling networks due to their compact formulation and ability to adequately represent network dynamics without the need for chemical kinetics. The problem posed in the SIGNET challenge is to determine the set of Boolean rules that describe the interactions of nodes within a plant signaling network, given a set of 300 Boolean pseudodynamic simulations of the true rules. The two solution methods that were presented revealed that the problem can be solved to greater than 99% accuracy.
Keywords: Boolean pseudodynamics, plant signaling network,

1 Introduction

Development of accurate models to predict cellular response to stimulus must begin with a proper characterization of the interaction between the various cellular processes. It is estimated that each individual gene or protein, on average, interacts with four to eight other genes and is involved in ten biological functions [M. I. Arnone *et al.*, 1997; G. L. Miklos *et al.*, 1996]. A seamless interaction between all cellular processes is essential for a living cell to thrive.

Kinetic models have been successfully applied to the analysis of a wide variety of biological systems, recent examples include neuronal signaling and the role of synaptic plasticity [S. M. Ajay *et al.*, 2006], phase sensitivity in circadian rhythms [R. Gunawan *et al.*, 2007], and prediction of IL-2 response from T-cell receptor activation [M. L. Kemp *et al.*, 2007]. By providing a global view of the underlying system, a kinetic model can be used to interpret new experimental data in the proper biological context [G. von Dassow *et al.*, 2000], provide mechanistic explanations for counter-intuitive observations [E. M. Fallon *et al.*, 2000], and facilitate the formulation of experimentally testable hypotheses [W. N. Abouhamad *et al.*, 1998; D. Endy *et al.*, 2000]. Unfortunately, accurate descriptions of underlying chemical kinetics are difficult to determine *in vivo*, with reliable kinetic coefficient estimation being a non-trivial and frequently impossible challenge due to a lack of identifiability [K. Yao *et al.*, 2003].

Experimental observations of cellular function indicate that the input-output behavior of signaling networks has a sigmoidal time dependence, and often can be adequately explained using the Heaviside, or step function [R. Thomas, 1973]. This observation suggests that a two state Boolean model could be employed to represent signaling network nodes, with nodal values being determined using an associated logical rule, representing network edges. Recent research has focused on applying rule-based Boolean models to the challenging problem of predicting biological network dynamics [S. Li *et al.*, 2006]. In a Boolean network model, the nodes of the network represent biological entities and the edges represent the interactions between them. The nodes can have a value of 0 or 1, representing an inactive or an active state, respectively. The

network dynamics are determined by Boolean rules for each node, that determine the state of the node at the next time-step based on the state of the upstream nodes, and the nodal update strategy. Rule-based Boolean network models have been successfully used to aid in explaining experimentally observed robustness of cellular networks [R. Albert *et al.*, 2003; S. Kauffman *et al.*, 2003; R. Thomas, 1973], and to determine the effects of an alteration in the network components and individual reaction rates [M. Chaves *et al.*, 2005].

At CFDRC, we have developed an augmented Boolean pseudo-dynamics approach to identify and quantitatively rank the importance of a node using a Boolean description of a cellular interaction network. The approach, known as the Boolean Network Dynamics and Target Identification (BNDTI), combines network topology and dynamic state information to determine the relative importance of a particular node with respect to the overall response of the network [A. S. Soni *et al.*, 2008]. In order to perform a demonstration of the utility of the newly developed approach, the guard cell signaling network in plant cells was selected [S. Li *et al.*, 2006]. This signaling network has been painstakingly translated into a Boolean network, and centers around abscisic acid (ABA) signal transduction, which for many decades has been known to play a role in ABA induced stomatal closure, regulating the plant water balance and imparting drought resistance. Two major secondary messengers involved in the closure of the stomata via ABA signal transduction are cytosolic calcium and the cytosolic pH. These two messengers are in turn regulated by a variety of other enzymes, secondary messengers, small molecules, and membrane channels. Figure 1 is a rendering of the interaction network, illustrating the complex regulatory interactions between species.

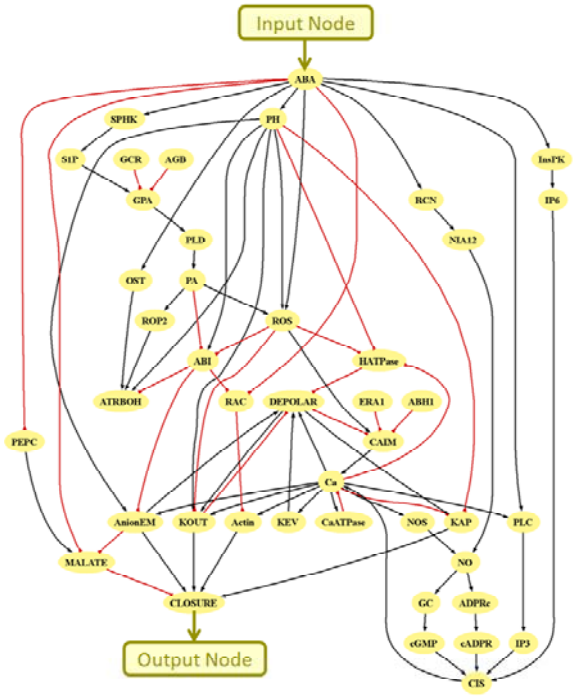

Figure 1: A Schematic of the Guard Cell Signaling Network. Inhibition reactions are shown with red edges and inverse arrowheads, whereas activation interactions are shown as black edges.

In the remainder of the paper, Section 2 provides a statement of the particular problem posed in the SIGNET challenge, along with some comments on the importance of the problem and how researchers addressed the problem. Section 3 includes a summary of the challenge results along with relevant comments.

2 SIGNET Challenge Problem Description

The problem posed in the SIGNET challenge is to determine the set of Boolean rules that describe the interactions of nodes within a plant signaling network, given a set of 300 Boolean pseudodynamic simulations of the true rules. The relevance of this problem arises from the trend in the biological sciences toward the increased availability of large datasets generated using high-throughput, high-content experimental technologies (such as gene expression microarrays). Experimental methods are currently able to probe the interactions of many thousands of cellular components simultaneously. However, cellular signaling pathways are still one of the most challenging and illusive types of networks to access experimentally due to the lack of methods for determining the state of a signaling network in an intact living cell. The SIGNET problem anticipates that experimental techniques for signaling network measurements will continue to progress, and assumes the availability of a large set of high throughput data that will be used to determine the set of Boolean rules describing the signaling network.

The expense and time in signaling measurements necessitate that the majority of signaling network models in the published literature are manually constructed using the relatively sparse literature data. The typical methodology includes a thorough assimilation of all relevant literature, followed by the construction of a table that formalizes the nodes (components) and edges (interactions) of the network. Using the table, a necessary and sufficient network capable of predicting the relevant behavior is generated. Often times the network is manually generated, introducing human bias and utilizing a significant amount of time and resources.

Automated methods for Boolean network inference have focused a significant amount of attention to the problem of identifying gene networks. The REVEAL (REVerse Engineering ALgorithm) was one of the first algorithms designed for this purpose [S. Liang *et al.*, 1997], which combines information-theory tools with an exhaustive search to generate a network that is consistent with the data. An alternative algorithm is the BOOL-1 algorithm [T. Akutsu *et al.*, 1999], which consists of examining all possible k-tuples of inputs and testing all Boolean function for each k-tuple until a consistent network is generated. The difficulty has motivated the utilization of heuristic approaches. An example of a heuristic approach is ID3 [J. Quinlan, 1986], which is a well known algorithm in Machine Learning. ID3 is based on the incremental construction of the input set for each variable using a greedy search. The approach presented in next section is based on the synergistic utilization of evolutionary algorithms and existing heuristics such as ID3. More recent approaches include the p-ary transitive reduction (TR_p) [R. Albert *et al.*, 2007], have been demonstrated that produce an optimal network given the constraints of minimal false positive inferences. Unfortunately, due to the lack of the necessary quantities of experimental data little effort has been expended for the automated identification of cellular signaling networks. Therefore, the overall goal of the SIGNET challenge was, therefore, to increase awareness of this problem area and stimulate interest in novel methods of solution.

The SIGNET dataset was generated using the procedure that follows. Nodes, edges, and Boolean rules were obtained from the work of Li, *et. al.* [S. Li *et al.*, 2006]. The network consists of a total of 43 nodes. Five nodes are input nodes (nodes that have only out-degree), and are the initiators of network action. The state of the input nodes ABA, GCR, ABH1, ERA1 was fixed at a value of '1' throughout the simulation. The state ('0' or '1') of the remaining 38 variable nodes was selected at random at the start of each simulation. 300 Boolean pseudodynamic (BPD) simulations were then generated using the asynchronous update strategy. The choice of BPD update scheme depends on the distribution of kinetic timescales within the network. The two most popular update schemes are synchronous and asynchronous. The synchronous method updates nodes in a fixed order at each time step, the order being determined at the start of the simulation. Synchronous updating assumes that the physical interactions within the network all occur at approximately the same time scale. Though synchronous updating is an efficient simulation method, it is rarely used for realistic systems due to the limiting assumption of similar time scales. In contrast, the asynchronous update method randomly determines the update order at each time step, which is equivalent to the assumption that the kinetic time scales within the network have a Gaussian distribution. Asynchronous updating is known to mimic realistic events in complex networks, and has been shown to effectively capture rare events [M. Chaves *et al.*, 2005; S. Li *et al.*, 2006]. Figure 2 is a plot of the response of the CLOSURE node averaged over the 300 randomly selected initial conditions with ABA=1 and ABA=0.

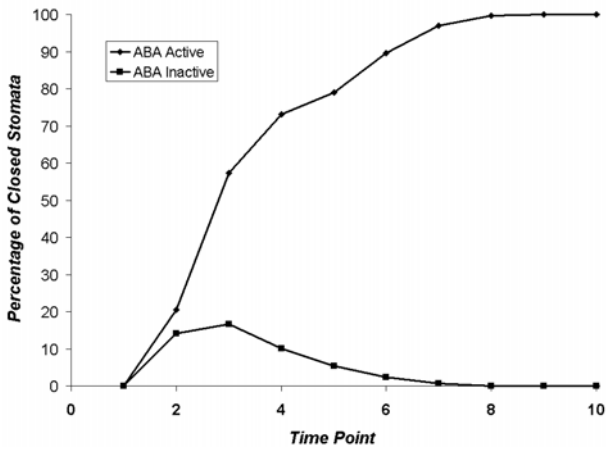

Figure 2: Effect of the presence and absence of abscisic acid on the percentage of closed stomata in the plant guard cell signaling model.

The overall objective of the SIGNET challenge was to determine the set of Boolean rules that describe the interactions of the nodes within this plant signaling network. The dataset includes 300 separate Boolean pseudodynamic simulations of the true rules, using an asynchronous update scheme. The results for 300 separate simulations are included in the dataset. Each simulation consists of a matrix of 0's and 1's, with 21 rows and 43 columns. The first row is the randomly generated initial condition for the particular simulation, with the next 20 rows being the output from the Boolean pseudodynamics simulation. Each of the 43 columns represents the transient response of a particular node. The nodal names are identified at the top of the data file.

3 Summary of SIGNET Challenge Results

Solutions to the SIGNET challenge were submitted by Mehreen Saeed of the Department of Computer Science at the National University of Computer and Emerging Sciences (Lahore Campus, Pakistan) [M. Saeed, 2009], and Cheng Zheng of the School of Mathematical Sciences at Peking University (Beijing, China) [C. Zheng *et al.*, 2009].

3.1 Performance Assessment

Solution methodology performance was assessed using the original SIGNET case, and for a second case generated by Prof. Isabelle Guyon. The organizers of the challenge provided a Matlab code for the evaluation of the algorithm performance. The evaluation code consisted of the generation of a truth table for each true rule and computing a prediction error by comparing

output values of the extracted rule to the output from the true rule. The error is computed by averaging over all rules. In addition to calculating the overall error rate of prediction, Prof. Saeed also calculated the training set error of the inferred rules. This was done by applying each rule to the individual Boolean vectors of the simulation data and predicting the output value. The output value was compared with the actual value to obtain the overall training accuracy rate.

3.2 Bernoulli Mixture Model (BMM)

The paper submitted by Prof. Saeed develops a Bernoulli distribution-based probabilistic model for the data, and combines this with the mixture densities to identify the Boolean rules from the SIGNET dataset. Parameters for the underlying Bernoulli distribution are estimated from the raw data using the expectation maximization (EM) algorithm. This methodology is stated to be ideal for estimating the probability distribution of non uni-modal data. Prof. Saeed has considerable experience applying this same methodology to the problem of dimensionality reduction and feature selection.

Optimal values for the number of mixtures as well as the probability thresholding value are given. The number of mixtures determines the underlying Bernoulli distribution, and the complexity of parameter extraction for estimating priors. Probability thresholding values are used to identify high data density areas on the corners of a hypercube. Each corner represents a conjunct of Boolean variables and together the set, of all the corners, forms a disjunction of rules, yielding a disjunctive normal form of a Boolean rule.

The results presented indicate that three mixtures produced the optimal training and evaluation accuracy of 94.55% and 82.98%, respectively, for the original SIGNET set. The dataset generated by Prof. Isabelle Guyon yielded a training accuracy of 95.88% and an evaluation accuracy of 87.61% for the three mixture model. A thresholding value of 0.70 produced good results for the case of a single mixture, but poor results for 2 and 3 mixtures. A thresholding value of 0.80 produced optimal results for 3 mixtures, and showed good accuracy for 1 and 2 mixtures. Thresholding values of 0.90 produced low accuracy results due to the number of results being ignored.

3.3 Minimum Explanatory Set and Maximum Likelihood (MESML)

The paper submitted by C. Zheng uses a method for finding the minimum explanatory set for a particular node [T. Ideker *et al.*, 2000], and then determines a Boolean function that generates maximal log likelihood for a particular node. The methodology is specifically modified for the reconstruction of asynchronous Boolean networks, where the nodal update order is selected at random.

Accuracy of the method was assessed in the same manner as with Prof. Saeed's solution. The accuracy of the proposed method was evaluated on the original SIGNET dataset and two other datasets generated by C. Zheng. Accuracy rates as a function of the number of assumed parent nodes are given for evaluation of the method. Interestingly, C. Zheng finds that as the number of parent nodes increases, the accuracy rate also decreases. This result is in agreement with the expectation that the average nodal accuracy should exhibit a maximum around the most probable in-degree, which for this network is 1 (58% of nodes). The averaged accuracy rate for a single parent node is 95%, which is excellent.

3.4 Discussion

The primary strength of the BMM methodology is the straight forward, novel approach of converting a probabilistic model into a rule based model in an intuitive manner. Information concerning the total runtime to expect in practice was not provided in the final manuscript, which would have aided the reader in making an implementation decision. However, the only bottleneck to performance would be the expectation maximization and I would not anticipate that it scales poorly with the number of mixtures.

The MESML method demonstrated by C. Zheng is the most accurate, with an average accuracy of 95%. The major drawback of the methodology is that the computational time scaling is roughly proportional to 10^n (see Table 2 of [C. Zheng *et al.*, 2009]), where n is the number of parent nodes. This is likely to cause a potential problem for networks that contain a large number of hub nodes, where the most probable in-degree is larger than one.

As experimental techniques become more sophisticated, computational methods will be called upon to provide biologically relevant insight into cellular behavior and interactions. Boolean networks will continue to play an ever increasing role in signaling network modeling due to their simplicity and predictive capability. Based upon the accuracy of the predictions, the results of the SIGNET challenge should provide significant confidence to researchers seeking to unravel the secrets of signaling networks.

Acknowledgements

I would like to thank Constantin Aliferis (NYU Medical Center) for encouraging our submission of the SIGNET dataset. I thank Prof. Saeed and C. Zheng for submitting their solutions. Guidance and encouragement provided by Isabelle Guyon and the organizers of the challenge is also acknowledged. Funding was provided by the U.S. Army MRMC (Program Manager: COL Alan Magill, WRAIR).

References

(1) W.N. Abouhamad, D. Bray, M. Schuster, K.C. Boesch, R.E. Silversmith, and R.B. Bourret, Computer-aided resolution of an experimental paradox in bacterial chemotaxis. J Bacteriol 180 (1998) 3757-64.

(2) S.M. Ajay, and U.S. Bhalla, Synaptic plasticity in vitro and in silico: insights into an intracellular signaling maze. Physiology (Bethesda) 21 (2006) 289-96.

(3) T. Akutsu, S. Miyano, and S. Kuhara, Identification of genetic networks from a small number of gene expression patterns under the boolean network model. Pacific Symposium on Biocomputing 4 (1999) 29.

(4) R. Albert, B. DasGupta, R. Dondi, S. Kachalo, E. Sontag, A. Zelikovsky, and K. Westbrooks, A novel method for signal transduction network inference from indirect experimental evidence. J Comput Biol 14 (2007) 927-49.

(5) R. Albert, and H. Othmer, The topology of the regulatory interactions predicts the expression pattern of the Drosophila segment polarity genes. J. Theor. Biol. 223 (2003) 1-18.

(6) M.I. Arnone, and E.H. Davidson, The hardwiring of development: organization and function of genomic regulatory systems. Development 124 (1997) 1851-64.

(7) M. Chaves, R. Albert, and E.D. Sontag, Robustness and fragility of Boolean models for genetic regulatory networks. J Theor Biol 235 (2005) 431-49.

(8) D. Endy, L. You, J. Yin, and I.J. Molineux, Computation, prediction, and experimental tests of fitness for bacteriophage T7 mutants with permuted genomes. Proc Natl Acad Sci U S A 97 (2000) 5375-80.

(9) E.M. Fallon, and D.A. Lauffenburger, Computational model for effects of ligand/receptor binding properties on interleukin-2 trafficking dynamics and T cell proliferation response. Biotechnol Prog 16 (2000) 905-16.

(10) R. Gunawan, and F.J. Doyle, 3rd, Phase sensitivity analysis of circadian rhythm entrainment. J Biol Rhythms 22 (2007) 180-94.

(11) T. Ideker, V. Thorsson, and R. Karp, Discovery of Regulatory Interactions Through Perturbation: Inference and Experimental Design. Pacific Symposium on Biocomputing 5 (2000) 302-313.

(12) S. Kauffman, C. Peterson, B. Samuelson, and C. Troein, Random Boolean network models and the yeast transcription network. Proc Natl Acad Sci USA 100 (2003) 14796-14799.
(13) M.L. Kemp, L. Wille, C.L. Lewis, L.B. Nicholson, and D.A. Lauffenburger, Quantitative network signal combinations downstream of TCR activation can predict IL-2 production response. J Immunol 178 (2007) 4984-92.
(14) S. Li, S. Assmann, and R. Albert, Predicting essential components of signal transduction networks: a dynamic model of guard cell abscisic acid signaling. PLoS Biol 4 (2006) e312.
(15) S. Liang, S. Fuhrman, and R. Somogyi, REVEAL, a general reverse engineering algorithm for inference of genetic network architectures. Pacific Symposium on Biocomputing 3 (1997) 29.
(16) G.L. Miklos, and G.M. Rubin, The role of the genome project in determining gene function: insights from model organisms. Cell 86 (1996) 521-9.
(17) J. Quinlan, Induction of decision trees. Machine Learning 1 (1986) 106.
(18) M. Saeed, The Use of Bernoulli Mixture Models for Identifying Corners of a Hypercube and Extracting Boolean Rules From Data. JMLR: Workshop and Conference Proceedings, this issue (2009).
(19) A.S. Soni, J.W. Jenkins, and S.S. Sundaram, Determination of critical network interactions: an augmented Boolean pseudo-dynamics approach. IET Syst Biol 2 (2008) 55-63.
(20) R. Thomas, Boolean formalization of genetic control circuits. J Theor Biol 42 (1973) 563-85.
(21) G. von Dassow, E. Meir, E.M. Munro, and G.M. Odell, The segment polarity network is a robust developmental module. Nature 406 (2000) 188-92.
(22) K. Yao, B. Shaw, B. Kou, K. McAuley, and D. Bacon, Modeling ethylene/butene copolymerization with multi-site catalysts: Parameter estimability and experimental design. Polym. React. Eng. 11 (2003) 563–588.
(23) C. Zheng, and Z. Geng, Reverse Engineering of Asynchronous Boolean Networks via Minimum Explanatory Set and Maximum Likelihood. JMLR: Workshop and Conference Proceedings, this issue (2009).

Appendix A: Pot-luck causality challenge: FACT SHEET (donated dataset)

Repository URL: http://www.causality.inf.ethz.ch/repository.php?id=5

Dataset name: SIGNET

Title: Boolean Rule Determination for Abscisic Acid Signaling

Author: Jerry W. Jenkins

Contact name, address, email and website: Jerry W. Jenkins
601 Genome Way, Suite 2301
Huntsville, AL 35806
http://www.cfdrc.com

Key facts:
Data dimensions (number of variables, number of entries), variable types, missing data, etc. See ftp://ftp.ics.uci.edu/pub/machine-learning-databases/DOC-REQUIREMENTS for inspiration.
Simulated data with a Boolean network modeling a biological signaling network.
Time series of 21 time steps. Initial step randomly drawn.
Number of variables: 43.
Number of entries: 300 separate dynamic simulations.
Variable types: binary.
Missing data: No.
During simulation, 38 of the 43 nodes are allowed to vary, with 5 nodes held constant throughout the simulation.

Abstract:
The objective is to determine the set of Boolean rules that describe the interactions of the nodes within this plant signaling network. The dataset includes 300 separate Boolean pseudodynamic simulations of the true rules, using an asynchronous update scheme. Each of the 300 simulations begin with a randomly generated initial condition, in order to ensure sampling of all of the steady states of the system. There are a total of 43 nodes in this dataset, with 5 nodes being constants.

The results for 300 separate simulations are included in the dataset. Each simulation consists of a matrix of 0's and 1's, with 21 rows and 43 columns. The first row is the randomly generated initial condition for the particular simulation, with the next 20 rows being the output from the Boolean pseudodynamics simulation. Each of the 43 columns represent the transient response of a particular node. The nodal names are identified at the top of the data file. A line of asterisks is used to separate the simulations from one another. An example set of data is included below:
****
10111011101011011011010010100011011000011001
11000011101111011011011111111011001011101011
11000111101111101011011000110100011110101010
11000011101111101011011000110000111101010 10
11000011101111101011011000110000111101010 10
11000011101111101011011000110000111101010 10
11000011101111101011011000110000111101010 10
11000011101111101011011000110000111101010 10
11000011101111101011011000110000111101010 10
11000011101111101011011000110000111101010 10
11000011101111101011011000110000111101010 10
11000011101111101011011000110000111101010 10
11000011101111101011011000110000111101010 10

11000011101111101011011000110000111101010101
11000011101111101011011000110000111101010101
11000011101111101011011000110000111101010101
11000011101111101011011000110000111101010101
11000011101111101011011000110000111101010101
11000011101111101011011000110000111101010101
11000011101111101011011000110000111101010101
11000011101111101011011000110000111101010101

Suggested task: Uncover the 43 Boolean rules x_i = f(x_1, x_2, ... x_43) of the Boolean Network.

We suggest to report results in disjunctive normal form (DNF), see, e.g., http://en.wikipedia.org/wiki/Disjunctive_normal_form, denoting the Boolean operators as "or", "and", and "not" and using regular parentheses.

Example:
ABI = (pH and not PA and not ROS) or (ABA and Ca)
One way to obtain these DNF formulae is to generate truth tables, then use a program like Minilog http://en.wikipedia.org/wiki/Minilog to generate the formula.

We now provide the truth values of the Boolean rules for self evaluation:

NO = NIA12 and NOS
PLC = ABA and Ca
CAIM = (ROS or not ERA1 or not ABH1) and not DEPOLAR
GPA = (S1P or not GCR) and AGB
ATRBOH = PH and OST and ROP2 and not ABI
HATPase = not ROS and not PH and not Ca
MALATE = PEPC and not ABA and not AnionEM
RAC = not ABA and not ABI
Actin = Ca or not RAC
ROS = ABA and PA and PH
ABI = PH and not PA and not ROS
KAP = (not PH and not Ca) and DEPOLAR
Ca = (CAIM or CIS) and not CaATPase
CIS = (cGMP and cADPR) or (IP3 and IP6)
AnionEM = ((Ca or PH) and not ABI) or (Ca and PH)
KOUT = (PH or not ROS or not NO) and DEPOLAR
DEPOLAR = KEV or AnionEM or not HATPase or not KOUT or Ca
CLOSURE = (KOUT or KAP) and AnionEM and Actin and not MALATE
ABA = 1
ABH1 = 1
AGB = 1
ERA1 = 1
GCR = 1
ADPRc = NO
CaATPase = Ca
cADPR = ADPRc
cGMP = GC
GC = NO

InsPK = ABA
IP3 = PLC
IP6 = InsPK
KEV = Ca
NIA12 = RCN
NOS = Ca
OST = ABA
PA = PLD
PEPC = not ABA
PH = ABA
PLD = GPA
RCN = ABA
ROP2 = PA
S1P = SPHK
SPHK = ABA

For evaluation, we suggest that, for each true generative rule, you generate the truth table, and compute the prediction error rate by comparing the predictions made by the rule of the proposed model to the target values. Then average the error rates over all rules. This measure does not respect the "natural" distribution of states, but this may be a feature rather than a bug because, for causal models, one wants to be robust against changes in distribution.

We provide some Matlab code to score the results and eventually generate new data (see http://www.causality.inf.ethz.ch/data/@signet.zip):

==> Usage for scoring:

s=read_rules(signet, 'your_submission_file.txt');
err=compare_rules(s);

Here is how it works:
- for each rule "zozo = some_boolean_expr(some_variables)"
* pool together the variables in the true rule for zozo and the propose rule
* create input vectors for all possible assignments of values to these variables
* apply the true rule to each input vector to get the target variables T
* apply the proposed rule to get the predicted Y
* Compute the error rate (fraction of disagreements between Y and T)
- average the error rates over all rules.

==> Usage for generating data:

dat=gene(signet, num, v_ini);

v_ini = initial state (43 binary values)
num = number of time steps
Returns a data matrix.

Keywords:
Boolean network, signaling network, time series