Learning a Probabilistic Boolean Network Model from Biological Pathways and Time-series Expression Data

Vardaan Pahuja[1], Ritwik Kumar Layek[2] and Pabitra Mitra[3]

Abstract— **The problem of inferring a stochastic model for gene regulatory networks is addressed here. The prior biological data includes biological pathways and time-series expression data. We propose a novel algorithm to use both of these data to construct a Probabilistic Boolean Network (PBN) which models the observed dynamics of genes with a high degree of precision. Our algorithm constructs a pathway tree and uses the time-series expression data to select an optimal level of tree, whose nodes are used to infer the PBN.**

I. Introduction

Modelling cellular interaction dynamics has been one of the important issues in systems biology [1]. A number of mathematical formulations have been proposed to model these genetic interactions, including Bayesian networks [2], linear models [3], and Boolean networks [4]. Based on a couple of limitations of BNs (e.g. limitation that BN is a deterministic model), a stochastic version of BNs, i.e. Probabilistic Boolean Networks (PBN) was proposed by Shmulevish et al. [5]. It incorporates uncertainty both in data and model selection.

The task of inferring gene regulatory networks from prior biological data is an ill-posed inverse problem, since multiple network realizations could explain the same biological phenomenon. The search space for potential regulatory genes and the boolean functions associated with them, increases exponentially with the number of genes in the network. Use of biological pathways to infer boolean networks was demonstrated in [6]. Restricted boolean networks are simplified boolean networks in which the regulatory relationships between genes is either activation (positive regulation to target gene) or inhibition (negative regulation to target gene). A three-rule method to construct a restricted Boolean network from time-series data was proposed by Higa et al. [7].

In our paper, we propose a novel algorithm which utilizes both biological pathway data and time-series expression data to construct a Probabilistic Boolean network to model the gene dynamics. Earlier approaches use a single form of biological data, which could be subjected to experimental bias. We overcome this limitation in our algorithm by using two different forms of data to infer the PBN.

The organization of the paper is as follows. Section II describes the mathematical model of Boolean networks, Probabilistic Boolean networks, and Restricted Boolean networks. The three-rule method for inferring regulatory relationships from time-series data is also discussed. In Section III, we describe our proposed algorithm for inferring the PBN from both types of data. Section IV evaluates the inferred PBN with the ground truth. The results are presented in Section V. Section VI infers a PBN model for Yeast Cell Cycle using our algorithm. Section VII concludes the paper with discussion of challenges involved and future scope of the problem.

II. Preliminaries

A. Boolean Networks

1) Introduction: A Boolean network *(BN) G(V,F)* on n genes is defined by a set of nodes/genes such that each node has a Boolean function assigned to it. Here F is the set of Boolean functions where,

$$F = \{f_1, f_2, ..., f_n\}, f_i : \{0,1\}^n \to \{0,1\}, \tag{1}$$

and V is the set of nodes, V= $\{v_1, v_2, ..., v_n\}$. The value v_i denotes the state of gene i, which can be either 0(off) or 1(on). The dynamics of BN can be expressed as,

$$v_i\left(t+1\right) = f_i\left(v_1\left(t\right), v_2\left(t\right), ..., v_n\left(t\right)\right) = f_i\left(\boldsymbol{v}\left(t\right)\right) \tag{2}$$

Here $\boldsymbol{v}\left(t\right)$ is called the Gene Activity Profile (GAP).

2) Restricted Boolean Networks: Restricted Boolean networks are simplified Boolean networks in which the regulatory relationships between genes obey the following convention: $a_{ij} = 1$ represents a positive regulation from gene x_j to x_i (activation); $a_{ij} = -1$ represents a negative regulation from gene x_j to x_i (inhibition); and $a_{ij} = 0$ means that x_j has no effect on x_i. The Boolean function $f_i\left(x_1, ..., x_{k_i}\right)$ is defined as [8]

$$x_i\left(t+1\right) = \begin{cases} 1, \quad if \displaystyle\sum_{j\in\{1,...,k_i\}} a_{ij}x_j\left(t\right) > 0 \\ 0, \quad if \displaystyle\sum_{j\in\{1,...,k_i\}} a_{ij}x_j\left(t\right) < 0 \\ x_i\left(t\right), if \displaystyle\sum_{j\in\{1,...,k_i\}} a_{ij}x_j\left(t\right) = 0 \end{cases} \tag{3}$$

B. Probabilistic Boolean Networks

1) Introduction: BN is a deterministic model. However, due to inherent uncertainity associated with a biological system, a stochastic model is more appropriate here [5]. Probabilistic Boolean Network is a stochastic version of BN in which more than one Boolean function can be assigned to a gene. Thus, for every node, there corresponds a set

$$F_i = \{{f_j}^{(i)}\}_{j=1,2,...,l(i)} \tag{4}$$

[1,2] Vardaan Pahuja and R.K. Layek are with Department of Electronics and Electrical Communication Engineering, IIT Kharagpur. vardaanpahuja@iitkgp.ac.in[1], ritwik@ece.iitkgp.ernet.in[2].

[3] Pabitra Mitra is with Department of Computer Science and Engineering, IIT Kharagpur, India-721302 pabitra@cse.iitkgp.ernet.in[3]

978-1-4577-0220-4/16/$31.00 ©2016 IEEE

where each ${f_j}^{(i)}$ is a possible predictor function for gene i and $l\left(i\right)$ is the number of possible functions for gene i. The probability of choosing the j^{th} predictor function for gene i is c_i^j. This implies that

$$\sum_{j=1}^{l_i} c_i^j = 1, 0 < c_i^j < 1, \text{for } i = 1, 2, ..., n \tag{5}$$

If we choose the j_i^{th} Boolean function for gene v_i, then the BN can be expressed as $BN_{j_1,j_2,...,j_n}$ where $j_i \in \{1, 2, ..., l_i\}$.The probability of choosing $BN_{j_1,j_2,...,j_n}$ is given by

$$P\{f_1 = f_1^{j_1}, f_2 = f_2^{j_2}, ..., f_n = f_n^{j_n}\} = \prod_{i=1}^{n} c_i^{j_i} = q_{j_1 j_2 \dots j_n} \tag{6}$$

2) Gene Influence in PBN: Different genes can have a varying degree of impact on the predictor function of a gene. The partial derivative of a Boolean function with respect to variable $x_j \left(1 \leq j \leq n\right)$ is defined as

$$\frac{\partial f}{\partial x_j} = f\left(x^{(j,0)}\right) \oplus f\left(x^{(j,1)}\right) \tag{7}$$

where $\oplus$ is modulo-2 addition operation.

The influence of the variable x_i on function f_i is the expectation of the partial derivative with respect to initial joint probability distribution $D\left(x\right), x \in \{0, 1\}^n$.

$$\begin{aligned} I_j\left(f\right) = E_D\left[\frac{\partial f}{\partial x_j}\right] &= \Pr\left\{\frac{\partial f}{\partial x_j} = 1\right\} \\ &= \Pr\left[f\left(x\right) \neq f\left(x^{(j)}\right)\right] \end{aligned} \tag{8}$$

where $x^{(j)}$ is same as x except that the j^{th} component is toggled.

C. Biological Pathways

The pathway segment $A \xrightarrow{t:a,b} B$ implies that if gene A assumes the value a, then gene B transitions to b in no more than t subsequent time-stamps [6]. A pathway is defined to be a sequence of pathway segments of the form $A \xrightarrow{t_1:a,b} B \xrightarrow{t_2:b,c} C$. A trajectory is a sequence of states $S_0 \rightarrow S_1 \rightarrow S_2 \rightarrow S_3 \rightarrow S_4$ resulting from network rules beginning at some initial state. These pathways represent *a priori* biological information. Our goal is to generate a PBN whose trajectories are consistent with the given set of biological pathways.

D. Inferring Regulatory relationships using time-series expression data

TABLE I
REGULATORY RELATIONSHIPS FOR ONE INPUT GENE

ID	$x_{j_1}(t)$	$x_i(t) \rightarrow x_i(t+1)$	a_{ij_1}
1	1	$0 \rightarrow 0$	-1
2	1	$0 \rightarrow 1$	1
3	1	$1 \rightarrow 0$	-1
4	1	$1 \rightarrow 1$	1

1) Three-rule method: A time-series observation can be treated as a trajectory (or random walk) of the state space of the network used to model a real biological system. The three-rule method proposed by Higa et al. [7] is to induce the constraints between genes from the small difference between two similar states and the difference between their next states. Given an m-point time series $S = \{S\left(1\right), S\left(2\right), ..., S\left(m\right)\}$ of gene expression profiles, where $S\left(t\right) \in \{0, 1\}^n$ for $t = 1, 2, ..., m$, the three rules are as follows:

Rule 1: Let $S(t{-}1)$, $S(t)$, and $S(t{+}1)$ be three consecutive states. If $S(t-1)$ and $S(t)$ differ by a single gene x_k, then for each gene x_i such that $x_i\left(t\right) \neq x_i\left(t+1\right)$, we have x_k directly regulates x_i; that is, $a_{ik} \neq 0$.

Rule 2: Only the active genes at time t can possibly regulate genes at time $t + 1$.

Rule 3: Given two similar states $S(t_1)$ and $S(t_2)$, the difference between $S\left(t_1 + 1\right)$ and $S\left(t_2 + 1\right)$ must result from the genes in their predecessors $S(t_1)$ and $S(t_2)$ that are expressed differently.

Rules 1 and 3 are also applicable to situations where $S(t{-}1)$ and $S(t)$ or $S(t_1)$ and $S(t_2)$ differ in more than one gene. Cyclically applying these rules to any two states may lead to a group of constraint inequalities between variables a_{ij}.

2) Constraint based analysis of regulatory relationships: Here, we analyze the constraint inequalities in equation (3) and use it to infer the regulatory relationships. The target gene can switch its state in four different combinations i.e. $0 \rightarrow 0$, $0 \rightarrow 1$, $1 \rightarrow 0$, and $1 \rightarrow 1$. Only the input genes which are active at time $(t - 1)$, contribute to the change of state at time t. Using equation (3), the following inequalities are true for different cases:

$$\begin{aligned} 0 \rightarrow 0 : &\quad \sum_{j \in \{1,...,k_i\}} a_{ij} x_j\left(t\right) \leq 0 \\ 0 \rightarrow 1 : &\quad \sum_{j \in \{1,...,k_i\}} a_{ij} x_j\left(t\right) > 0 \\ 1 \rightarrow 0 : &\quad \sum_{j \in \{1,...,k_i\}} a_{ij} x_j\left(t\right) < 0 \\ 1 \rightarrow 1 : &\quad \sum_{j \in \{1,...,k_i\}} a_{ij} x_j\left(t\right) \geq 0 \end{aligned} \tag{9}$$

For a single regulatory gene x_{j_1}, these inequalities simplify to $a_{ij_1} = -1$, $a_{ij_1} = 1$, $a_{ij_1} = -1$, and $a_{ij_1} = 1$ respectively. These are presented in Table I. For the case of two regulatory genes, if a single gene is active, then the regulation of the active gene can be inferred but that of the other gene is undetermined. When both of input genes are active, the regulation of both these genes can be determined if the target gene switches its state. In the other case, the relationship between their regulatory relationships is semi-determined because it is governed by a constraint equation. The different cases for two input gene regulatory

TABLE II
REGULATORY RELATIONSHIPS FOR TWO INPUT GENES

ID	$x_{j_1}(t)$	$x_{j_2}(t)$	$x_i(t) \to x_i(t+1)$	a_{ij_1}	a_{ij_2}	Constraint
1	0	1	$0 \to 0$	No	-1	
2	1	0		-1	No	
3	1	1		-1 or 1	-1 or 1	$a_{ij_1}+a_{ij_2} \leq 0$
4	0	1	$0 \to 1$	No	1	
5	1	0		1	No	
6	1	1		1	1	
7	0	1	$1 \to 0$	No	-1	
8	1	0		-1	No	
9	1	1		-1	-1	
10	0	1	$1 \to 1$	No	1	
11	1	0		1	No	
12	1	1		-1 or 1	-1 or 1	$a_{ij_1}+a_{ij_2} \geq 0$

No: Undetermined ; -1 or 1: Semi-determined

relationships are presented in Table II. Similar constraint inequalities can be derived for three input gene regulatory relationships.

Each time series sample gives rise to one of the cases mentioned in the respected tables. Let N_{ij}^{-1}, N_{ij}^{1}, and $N_{ij}^{-1,1}$ denote the number of $a_{ij} = -1$, $a_{ij} = 1$, and $a_{ij} = -1$ or 1 respectively. The degree of determination of a regulatory relationship a_{ij} is defined as

$$d_{ij} = |N_{ij}^{-1} - N_{ij}^{1}| \tag{10}$$

Among multiple input genes in a regulatory relationship, the one with the highest d_{ij} is the first to be decided using majority rule. This value is then put into constraint inequalities for inferring other semi-determined relationships. This procedure is then repeated to determine all other regulatory relationships.

3) Error Analysis: The error arising out of ambiguity in determination of a_{ij} is defined as $\varepsilon_{ij}^{-1,1} = min\left(N_{ij}^{-1}, N_{ij}^{1}\right)$. Also, the target gene can't switch its state under null input conditions. This error is denoted by ε_{i}^{null}. The total error of a predictor set is defined as

$$\varepsilon = \varepsilon_{i}^{null} + \sum_{j} \varepsilon_{ij}^{-1,1} \tag{11}$$

4) Inference Algorithm: The algorithm used for determining regulatory relationships [9] is given below:

1. Calculate the total error of each combination of one, two, or three regulatory gene sets.
2. Sort the predictor sets in ascending order of their errors.
3. If a gene appears in the first l sets with a frequency greater than or equal to 50%, then it is selected as a regulatory gene.

III. PROPOSED ALGORITHM

A. Introduction

The list of biological pathways satisfied by the biological system is available to us. Further, we assume a priority ordering of these pathways in order of decreasing reliability i.e. pathways higher in order are more accurate than those lower in order. The complete set of pathways can't represent a Boolean network as many these pathways may conflict with each other regarding prediction of a gene output. On the other hand, a subset of pathways represents a family of Boolean networks since the Karnaugh maps representing the BNs can contain several don't care terms. Our proposed algorithm constructs a m-ary tree with each node containing a subset of non-conflicting pathways. The invariant followed is that each child node satisfies the pathways satisfied by the parent node i.e. the pathway set of a child node is a super-set of that of parent node.

B. Construction of Pathway Tree

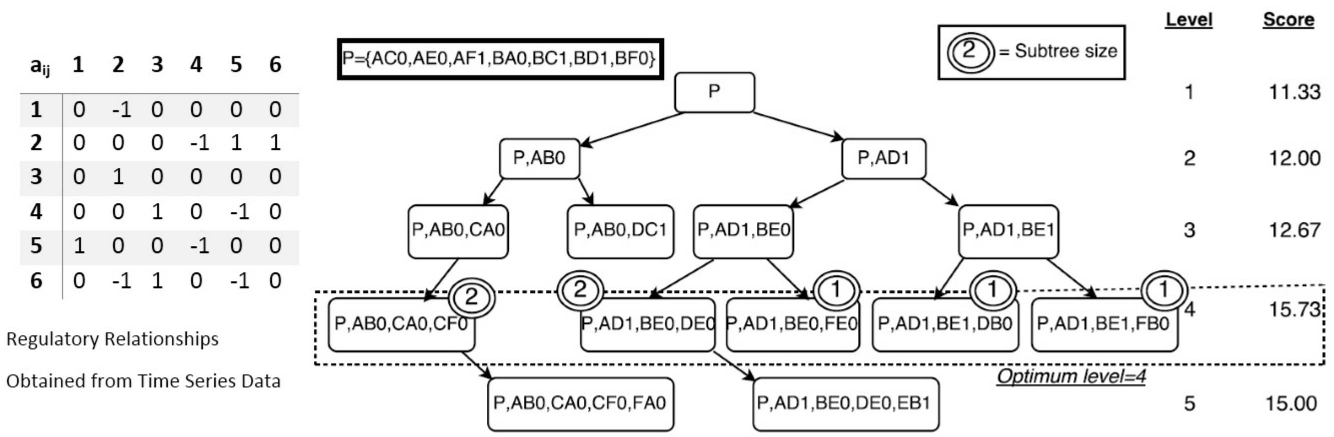
Here, we describe the method of constructing the pathway tree. The conflict of a pathway with a node indicates conflict with its pathway set. The following steps are to be followed for construction of tree:

1. The initial contiguous set of non-conflicting pathways is added to the root node.
2. For each new pathway in the list, traverse the tree from root till it gets added to a node's list of pathways. Three cases arise here:
   i) If the node has two children with only one of them conflicting with the current pathway, set the non-conflicting node as current node. If both children are non-conflicting, choose either one with equal probability. Else, create a new node containing the parent's list of pathways and add the current pathway to its list and stop.
   ii) If the node has only one child, and the child is conflicting, then proceed with creating a new node as mentioned earlier. Else, either choose the child as current node or create a new node(as mentioned earlier) with equal probability.
   iii) If the node has no children, then create a new node(similar to previous step).

   If a new node is created, the procedure terminates for the current pathway. Otherwise, steps (i), (ii) and (iii) are repeated for the new current node.

C. Selecting the optimum level of tree

Each level of the tree created, contains nodes representing a family of BNs and thus a PBN. The nodes in levels near the root contain fewer pathways while the ones at leaves have more pathways. Our goal is to strike a balance between them such that an optimum number of pathways, highly reliable according to information from time-series expression data are considered in our model. The time-series expression data gives us the regulatory genes(and their regulation: activation or inhibition) for each gene. At each level, we compute a score by summing the influences of these regulatory genes on BN across all target genes and across all nodes of that level. The score is then normalized by the number of nodes in that level. Let the number of nodes in i^{th} level be r_i



a_{ij}	1	2	3	4	5	6
1	0	-1	0	0	0	0
2	0	0	0	-1	1	1
3	0	1	0	0	0	0
4	0	0	1	0	-1	0
5	1	0	0	-1	0	0
6	0	-1	1	0	-1	0

Regulatory Relationships

Obtained from Time Series Data

Fig. 1. Pathway Tree and Regulatory Relationships for $n = 6, m = 10$

and let n_{ik} denote the k^{th} node in i^{th} level. Let S^i_{kp} denote the set of regulatory genes for p^{th} gene in n_{ik}. $I_j\left(f^{(p)}_{n_{ik}}\right)$ denotes the influence of gene x_j on the predictor function of p^{th} gene in n_{ik}. The score for level i is defined as,

$$\text{Score}\,(i) = \frac{\sum\limits_{k=1}^{r_i}\sum\limits_{p=1}^{n}\sum\limits_{j|x_j\in S^i_{kp}} I_j\left(f^{(p)}_{n_{ik}}\right)}{r_i} \tag{12}$$

The level with the highest score is selected as the optimum level, since it correlates the best with information from time-series expression data.

Thus, the PBN is constructed by using a linear combination of BN families of the optimum level and proportionally weighing each node by the size of its sub-tree.

IV. Performance Evaluation

In a n gene biological system, we randomly generate a set **P** of non-conflicting pathways. Then we create m sets of pathways, each containing the set **P** as subset, plus some additional pathways, non-conflicting with **P**. Now, each of m sets of pathways represents a BN family. We mix them in a random proportion to generate our ground-truth PBN. Let p be the vector of coefficients of these m BNs in the PBN.

Selecting a random initial state and performing Monte-Carlo simulations of the transition probability matrix of PBN gives us a time-series data of gene states (boolean values), which is then used to infer regulatory relationships. The pathway set used for constructing the tree contains the set **P** followed by other pathways in those m sets. The algorithm constructs a tree, with h^{th} level being optimal. Let us assume the h^{th} level has r nodes. The normalized hamming distance metric for comparing two BN families is

$$\mu_{ham} = \frac{1}{n*2^n}\sum_{i=1}^{n}\sum_{k=1}^{2^n}\left[f_i\left(x_k\right)\oplus f_i^{'}\left(x_k\right)\right], \tag{13}$$

where $f_i\left(.\right)$ and $f_i^{'}\left(.\right)$ represent Boolean functions of gene i in ground-truth and the inferred network ; x_k represents a binary state vector. The $\oplus$ operator returns 0.5 in case either of the operand is a don't care, while the usual definition holds for other cases. However, in our case, the ground-truth network contains a weighted combination of m BN families. Let the reconstructed network contains r BN families. Accordingly, the distance metric for comparison is as follows

$$\mu_{ham}^{'} = \frac{\sum\limits_{k=1}^{r}\left[\left(\sum\limits_{i=1}^{m}\mu_{ham}\left(n_k, BN_i\right)*p\left(i\right)\right)*size\left(n_k\right)\right]}{\sum\limits_{k=1}^{r}size\left(n_k\right)}, \tag{14}$$

where n_k denotes the k^{th} node in the optimum level h.

V. Results and Discussion

The experiment described in the previous section is performed for $m = 5, 10, 15$ and $n = 4, 5, 6, 7$ genes. For each case, the measure is averaged over an ensemble of 100 biological systems, each containing a unique set of pathways and time-series expression data, shown in Table III. The time-series contains 100 points and the value $l = 7$ was chosen for the inference algorithm. A sample output of the algorithm for $n = 6, m = 10$ including the tree, scores of different levels and the optimum level is shown in Fig. 1. Here, 'ABb' denotes the pathway $A \xrightarrow{1:1,b} B$ and **P** is the set of non-conflicting pathways. The output PBN is obtained by switching between $BN_1, ..., BN_5$ of optimum level 4 with probabilities $\left[\frac{2}{7}, \frac{2}{7}, \frac{1}{7}, \frac{1}{7}, \frac{1}{7}\right]$ respectively.

The distance between ground-truth and the reconstructed PBN reduces with increase in number of genes. This is due to relatively more number of pathways in networks with lesser number of genes. Thus, less number of constraints leads to more efficient reconstruction. Increase in m also reduces the distance measure. This shows more amount of data results in better inference. However, the difference between $m = 10$ and $m = 15$ is less pronounced for $n = 6, 7$. This could be due to saturation in quality of new data available. In fact, presence of more inconsequential pathway information can only degrade the accuracy of the model.

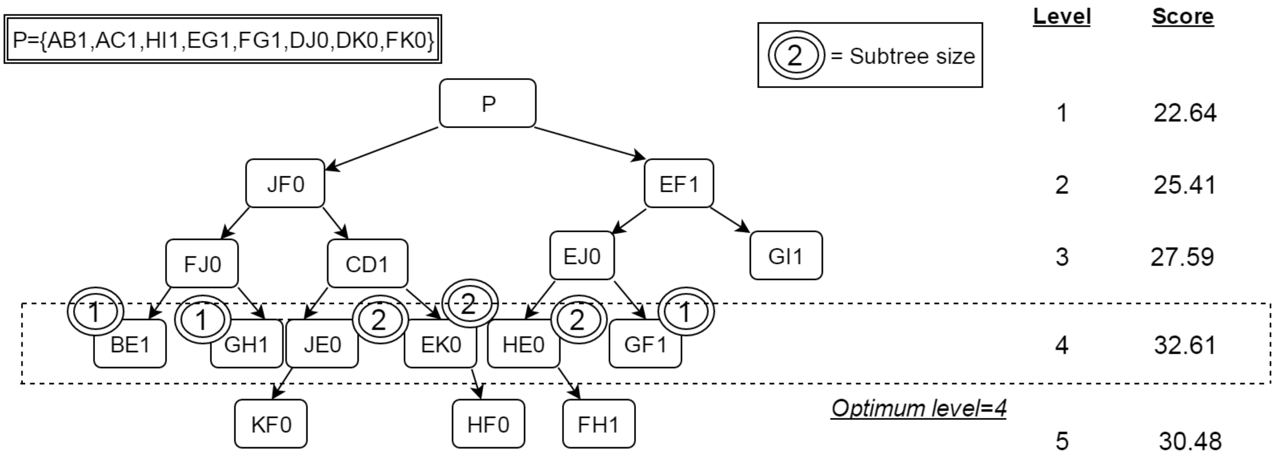


Fig. 2. Pathway Tree for Yeast Cell Cycle

TABLE III
DISTANCE MEASURE BETWEEN THE GROUND-TRUTH AND RECONSTRUCTED PBN

n	$m=5$	$m=10$	$m=15$
	$\mu_{ham}^{'}$	$\mu_{ham}^{'}$	$\mu_{ham}^{'}$
4	0.524	0.458	0.428
5	0.444	0.371	0.357
6	0.344	0.281	0.278
7	0.275	0.255	0.243

VI. INFERENCE OF YEAST CELL CYCLE NETWORK

The cell cycle is a vital biological process in which one cell grows and divides into two daughter cells. It consists of four phases, G1, S, G2, and M. Its regulation is highly conserved among eukaryotes [10]. From the 800 genes involved in cell cycle process of a budding yeast, Li et al. [8] constructed a network of 11 key regulators Cln3, MBF, SBF, Cln1, Clb5, Clb1, Mcm1, Cdc20, Swi5, Sic1, and Cdh1 which we shall refer to as A, B, C, D, E, F, G, H, I, J, and K respectively. We use the pathways and time-series expression data as given in [8] and compute the PBN using our algorithm. The resultant pathway tree is shown in Fig. 2. The output PBN switches between $BN_1, BN_2, BN_3, BN_4, BN_5$, and BN_6 of optimum level 4 with probabilities $[\frac{1}{9}, \frac{1}{9}, \frac{2}{9}, \frac{2}{9}, \frac{2}{9}, \frac{1}{9}]$ respectively. For convenience, only pathways distinct with parent node are shown for each node. Our algorithm incorporates the essential pathways and gives proportional weightage to other pathways in modelling the cell cycle trajectory.

The root of pathway tree represents pathways which are most essential for performing the cell cycle process. The successive higher levels of tree represent the decreasing level of significance in representing the cell cycle process. The set of pathways in the optimum level correlates best to the given time-series data.

VII. CONCLUSION

In this paper, we proposed an algorithm to infer a PBN for a biological system using biological pathways and time-series gene expression data. Pathways represent prior biological knowledge while time series data is obtained experimentally. The model space of PBN is huge compared to the amount of data available. Thus, a unique solution is impractical, given the fact that the data is noisy. Our solution overcomes this limitation by learning a model which makes systematic use of these two different forms of biological data. Future work will involve integration of multiple forms of such biological data to infer a more robust model.

REFERENCES

[1] S. Huang, "Gene expression profiling, genetic networks, and cellular states: an integrating concept for tumorigenesis and drug discovery," *Journal of Molecular Medicine*, vol. 77, no. 6, pp. 469–480, 1999.

[2] K. Murphy, S. Mian *et al.*, "Modelling gene expression data using dynamic bayesian networks," Technical report, Computer Science Division, University of California, Berkeley, CA, Tech. Rep., 1999.

[3] E. P. van Someren, L. F. Wessels, and M. J. Reinders, "Linear modeling of genetic networks from experimental data." in *ISMB*, 2000, pp. 355–366.

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

[5] I. Shmulevich, E. R. Dougherty, S. Kim, and W. Zhang, "Probabilistic boolean networks: a rule-based uncertainty model for gene regulatory networks," *Bioinformatics*, vol. 18, no. 2, pp. 261–274, 2002.

[6] R. K. Layek, A. Datta, and E. R. Dougherty, "From biological pathways to regulatory networks," *Molecular BioSystems*, vol. 7, no. 3, pp. 843–851, 2011.

[7] C. H. Higa, V. H. Louzada, T. P. Andrade, and R. F. Hashimoto, "Constraint-based analysis of gene interactions using restricted boolean networks and time-series data," in *BMC proceedings*, vol. 5, no. 2. BioMed Central, 2011, p. 1.

[8] 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.

[9] H. Ouyang, J. Fang, L. Shen, E. R. Dougherty, and W. Liu, "Learning restricted boolean network model by time-series data." *EURASIP J. Bioinformatics and Systems Biology*, vol. 2014, p. 10, 2014.

[10] A. W. Murray and T. Hunt, *The cell cycle: An Introduction*. Oxford University Press New York, 1993, vol. 251.