

# Influence of Prior Information on the Reconstruction of the Yeast Cell Cycle from Microarray Data

Frank Emmert-Streib

Stowers Institute for Medical Research  
1000 E. 50th Street  
Kansas City, MO 64110, USA  
Email: fes@stowers-institute.org

Matthias Dehmer

Max F. Perutz Laboratories  
Center for Integrative Bioinformatics Vienna  
Dr. Bohr Gasse 9  
A-1030 Vienna, Austria  
Email: matthias@dehmer.org

Chris Seidel

Stowers Institute for Medical Research  
1000 E. 50th Street  
Kansas City, MO 64110, USA  
Email: cws@stowers-institute.org

**Abstract**—In this paper we suggest a method to reconstruct the gene interaction network of the cell cycle of *S.cerevisiae* from microarray data. We study a small subnetwork comprising 20 genes and estimate iteratively partial correlations between expression profiles of the corresponding genes. Starting from a fully connected network an edge is deleted if the estimated partial correlation coefficient is not statistically significant from zero. To find an appropriate significance level for the statistical test we use prior information from the literature and calculate a so called efficient significance level used for the hypothesis tests. That means, our method is supervised and by this adaptive to various noise levels in the data.

## I. INTRODUCTION

In molecular biology the advent of high-throughput technologies have opened the possibility to record information on a genomic-scale. Theoretically, application of appropriate data analysis methods to these data should allow to reconstruct the molecular information flow, at least to a certain extend, to unravel the function of subprocesses and components maintaining, e.g., the cells ability to mitosis or apoptosis. The first generation of data analysis methods that have been applied to, e.g., data from microarray experiments, were clustering methods [5]. The goal of the data clustering was to associate, e.g., genes with known function with genes with unknown function. In this way, valuable information could be gained at a high rate ranging from cells of lower to higher organisms. On a downside, clustering methods provide only an unordered set allowing a rough association between the set members. However, no structural information between the set members can be inferred. Depending on the biological context such structural information could clarify questions concerning, e.g., transcription regulation, signaling or metabolism and, hence, provide insights in questions concerning the functional organization within a cell. In general, the functional units of the types mentioned above are summarized under the term *gene networks*, because a directed graph allows to visualize the information flow conveniently. The important question arising now is how to reconstruct the network structure of gene networks from high-throughput data. In this paper we investigate this question.

JUDEA PEARL et al. and PETER SPIRITES et al. suggested independently a mathematical framework based on the notion

of d-separation and higher-order partial correlations to infer causal structures, that means directed graphs, from data. More precisely, in [15] the PC-algorithm was introduced which iteratively estimates partial correlation coefficients whose order increases successively during the iterations. Starting with a fully connected graph between all variables, partial correlation coefficients are estimated from experimental data starting from low going to high orders. If one of these tests gives a vanishing partial correlation coefficient the edge between the corresponding variables is deleted, because a direct interaction could be ruled out. Theoretically, the PC-algorithm is able to reconstruct the network structure from given experimental data supposed all partial correlations of arbitrary order can be estimated reliably. However, practically the PC-algorithm is not applicable if the number of variables is high because of the combinatorial explosion of possible test that need to be calculated. For this reason, it was speculated that already low order partial correlations are sufficient to reconstruct the underlying network structure with high precision [4]. However, so far these approaches are solely based on heuristic considerations rather than a solid mathematical justification.

In this paper we introduce an algorithm similar to the PC-algorithm but with two significant differences. First, we restrict the maximal order of the partial correlation coefficients to three. This is one order higher compared to the approach used by [4] and allows to study the influence of the order on the number of detected edges. Second, we learn an efficient significance level from the data based on the assumption that we already know from the literature that some gene interactions (edges) should be present. That means, we correct values of the significance level by comprising the variability within the data quantifiable, because prior information about the biological system, i.e., that network structure, is available from the literature. This results in a supervised rather than unsupervised method as the PC-algorithm.

This paper is organized as follows. Starting from an introduction in section I, section II provides a short description of the genes participating in the yeast cell cycle we use in our study. Section III presents the main approach for reconstructing the network structure from microarray data. The experimental section IV summarizes the numerical results. The

paper finishes in section V with conclusions and a summary.

## II. THE YEAST CELL-CYCLE

In our study we consider only a small subnetwork of the yeast cell cycle consisting of 20 genes given in table I. In the following we give a short description of these genes and their role during the cell cycle.

Cyclins are regulatory subunits for Cyclin Dependent Kinases (CDKs). Central to the cell cycle in yeast is the protein kinase Cdc28, a CDK which serves to regulate many aspects of the cell cycle. Activation of CDKs occurs by cyclin binding, followed by stimulatory phosphorylation. Activation can be inhibited by the binding of inhibitory proteins, or through inhibitory phosphorylation. CLN1, CLN2, and CLN3 are G1 cyclins which promote the transition from G1 to S. Transcription of CLN1 and CLN2 is cell cycle dependent. They are expressed in G1, and their expression is dependent on the transcription factor complexes MBF and SBF. MBF consists of SWI6 and MBP1, while SBF consists of SWI6 and SWI4. CLN3 transcription occurs throughout the cell cycle. However the protein is regulated post-translationally through several PEST motifs. CLN3 is phosphorylated by Cdc28p, leading to its degradation. CLN3 is involved in the regulation of CLN1 and CLN2. CLB1, CLB2, CLB3, CLB4, CLB5, CLB6 are B-type cyclins which activate Cdc28 at various points in the cell cycle including S, G2, and M. CLB1 and CLB2 are expressed in G2 and promote the transition from G2 to M. They are both transcriptionally and post-translationally regulated. CLB1 is important for meiosis, while CLB2 is involved only in mitosis. CLB3 and CLB4 are expressed in S phase and also involved promoting G2 to M. CLB5 and CLB6 are important for initiation of DNA synthesis (initiation of S-phase). CLN1 and CLN2 are regulated by MBP1, a DNA binding protein that is part of the MBF complex which regulates genes involved in the G1/S transition of the cell cycle. The MBF complex consists of MBP1 and SWI6. CLN1 and CLN2 are also regulated by a related complex termed SBF, which consists of SWI6 and SWI4. SWI4 is highly homologous to MBP1. SWI5 is a transcription factor involved in expressing genes required for the M/G1 transition. CDC20 is an activator of the Anaphase Promoting Complex (APC), which controls the transition between metaphase and anaphase, and also controls exit from mitosis and entry into G1. Cdc20 is important for ubiquitin mediated degradation of CLB5 and CLB3 proteins via the APC. Cdc20 is cell cycle regulated, increasing expression as cells enter mitosis, and decreasing expression as cells exit mitosis. HCT1 (CDH1) is also an activator of the APC, and effects substrate specificity. HCT1 is required for APC mediated degradation of the CLB2 protein. HCT1 also interacts with CLB3. CDC34 is an E2 ubiquitin conjugating enzyme which, along with SKP1 is part of a complex involved in protein degradation. The SCF complex promotes the transition from G1 to S by targeting degradation of G1 cyclins, and SIC1 a cyclin/CDK inhibitor. MCM1 is a transcription factor involved in transcription of mating type specific genes (i.e. activation of alpha-specific genes,

TABLE I  
GENE INTERACTIONS FOUND IN THE LITERATURE. THE INTERACTION PARTNERS FORM ANY KIND OF PHYSICAL INTERACTION WITH A GENE OR PRODUCTS THEREOF.

Gene	Interaction partners	References from the literature
CLN1		
CLN2	CLB1, CLB2	[1], [17]
CLN3		
CLB1	SWI5, CLB2, CLN2	[1], [1], [1]
CLB2	CLB1, CLB6, CLN2, SWI5	[1], [1], [17], [9]
CLB4		
CLB5	CDC20	[8]
CLB6	CLB1, CLB2, CDC20	[1], [1], [8]
MCM1	CLB1, CLB2, SWI5	[1], [1], [1]
SIC1	CLB1, CLB2, SWI5	[17], [17], [17]
SWI6	SWI4	[8]
CDC28	SWI4, SWI6, CDC20	[8], [8], [8]
CDC53		
MBP1	CDC34, SKP1	[7], [2]
CDC34	MBP1	[7]
SWI5	CLN1, CLB1, CLB2	[1], [1], [9]
SKP1	MBP1	[2]
SWI4	CDC28, SWI6	[8], [8]
CDC20	CLB5, CLB6, CDC28	[8]
HCT1		

repression of  $\alpha$ -specific genes). Also found to affect regulation of genes involved in the cell cycle, such as CLN1, CLN2, CLB1, CLB2, SWI5.

## III. RECONSTRUCTION OF NETWORKS BASED ON PARTIAL CORRELATIONS OF HIGHER-ORDER

The algorithm we suggest to reconstruct the network structure from microarray data is similar to the PC-algorithm [15] with two significant differences. First, we restrict the order of the partial correlation to three. This is necessary, because for practical reasons it is not possible to estimate partial correlations up to an arbitrary order. A similar approach has been suggested by [4]. However, in this study partial correlations are limited to second-order. Second, due to the fact that we apply our method to biological data rather than to data from simulation studies we face the problem reliably estimating the partial correlations from the data. This means, that it is possible that a statistical test suggests to remove an edge between gene A and B, however, from the literature it is known that gene A and B interact biologically. If we assume, that the microarray experiment is thoroughly designed to capture this information we conclude that in this case the signal is rather weak but present and could just not pass the statistical test. To prevent such miss-judgments we suggest to adapt parameters of the statistical test used to distinguish vanishing from non-vanishing partial correlations from prior knowledge in the literature.

The central result we base on our investigations in this paper is from JUDEA PEARL et al. [18], [10]. They demonstrated that there is a correspondence between a graph theoretical entity called d-separation and the partial correlation coefficients of higher-order which will be repeated here for completeness.

*Definition 1:* (*d-separation*, [11]) A path  $p$  is d-separated (or blocked) by a set  $S$  iff one of the following statements holds:

- 1)  $p$  contains at least one non-collider  $z$  with  $z \in S$  **or**
- 2)  $p$  contains at least one collider  $z$  with neither  $z \in S$  nor any descendent  $z'$  of  $z$  is in  $S$

If a set  $S$  d-separates every path connecting  $x$  and  $y$  then  $S$  d-separates  $x$  and  $y$ ;  $x \perp\!\!\!\perp y | S$

*Theorem 1:* ([18], [6]) If the sets  $X$  and  $Y$  are d-separated by  $S$  in a DAG  $G$ , then  $X$  is independent of  $Y$  conditioned on  $Z$  in every Markovian model structured according to  $G$ . Conversely, if  $X$  and  $Y$  are not d-separated by  $S$  in  $G$ , then  $X$  and  $Y$  are dependent conditioned on  $Z$  in almost all Markovian models structured according to  $G$ .

Due to the fact, that conditional independence implies vanishing partial correlation coefficient theorem 1 can be summarized as

$$x \perp\!\!\!\perp y | S \iff \rho_{xy.S} = 0 \quad (1)$$

It has been shown by SPIRITES et al. [16] that Eq. 1 holds also for cyclic graphs if a linear model is assumed.

#### A. Undirected dependency graph

In the following we neglect the direction of an edge and reconstruct only undirected graphs from the data. This simplifies our approach and helps to focus on the major problem. We call the reconstructed graph *undirected dependency graph* (UDG) [13]. The UDG is obtained via the following algorithm 1. Here  $Pa(x)$  denoted the parent set of  $x$  and  $Z$  is a set of size  $O$ , e.g.,  $Z = \{x, y, z\}$  with  $x, y, z \in \{x\}$  for  $O = 3$ . For our numerical investigations we used  $N = 20$  genes listed in table I and  $O = 3$ .

*Definition 2 (UDG of third order):* An UDG  $G$  of third order is an undirected, unweighted graph with  $N$  nodes (number of genes) that is obtained via algorithm 1.

For the correlation we use PEARSON correlation

$$r_{\mathbf{x}\mathbf{y}} = \frac{C_{\mathbf{x}\mathbf{y}}}{\sigma_{\mathbf{x}}\sigma_{\mathbf{y}}} \quad (2)$$

with the covariance  $C_{\mathbf{x}\mathbf{y}} = E[(\mathbf{x} - \mu_{\mathbf{x}})(\mathbf{y} - \mu_{\mathbf{y}})]$ . The partial PEARSON correlation of first, second and third order are recursively given by

$$r_{xy.z_1} = \frac{r_{xy} - r_{xz_1}r_{yz_1}}{\sqrt{(1 - r_{xz_1}^2)(1 - r_{yz_1}^2)}} \quad (3)$$

$$r_{xy.z_1z_2} = \frac{r_{xy.z_1} - r_{xz_2.z_1}r_{yz_2.z_1}}{\sqrt{(1 - r_{xz_2.z_1}^2)(1 - r_{yz_2.z_1}^2)}} \quad (4)$$

$$r_{xy.z_1z_2z_3} = \frac{r_{xy.z_1z_2} - r_{xz_3.z_1z_2}r_{yz_3.z_1z_2}}{\sqrt{(1 - r_{xz_3.z_1z_2}^2)(1 - r_{yz_3.z_1z_2}^2)}} \quad (5)$$

Statistically, we test for vanishing (partial) correlation  $r$  by transforming  $r$  to

$$t_r = \frac{r\sqrt{N-2-O}}{\sqrt{1-r^2}} \quad (6)$$

---

#### Algorithm 1 Undirected dependency graph of third order

---

```

1: given  $N$  expression profiles  $\{\mathbf{x}\}$ 
2: each  $x$  is represented by one node
3: connect all  $N$  nodes with an edge
4: for all 2-tuples  $x$  and  $y$  from  $\{\mathbf{x}\}$  do
5:   estimate  $r_{xy}$ 
6:   if  $r_{xy} = 0$  then
7:     delete edge between  $x$  and  $y$ 
8:   end if
9: end for
10: for  $i = 1$  to  $O$  do
11:   for all  $x$  and  $y$  do
12:     determine  $Pa(x, y) = \{z | z \in Pa(x) \vee z \in Pa(y)\}$ 
13:     for all  $O$ -tuples  $Z$  from  $Pa(x, y)$  do
14:       estimate  $r_{xy.Z}$ 
15:       if  $r_{xy.Z} = 0$  then
16:         delete edge between  $x$  and  $y$ 
17:       end if
18:     end for
19:   end for
20: end for

```

---

The  $t_r$  values follow a Student's t distribution with  $df = N - 2 - O$  degrees of freedom [12], [13]. The null hypothesis  $H_0 : r = 0$  is rejected if  $t_r$  is greater or equal than the tabled critical two-tailed value  $t_\alpha$  for a significance level  $\alpha$  by assuming a nondirectional alternative hypothesis  $H_1 : r \neq 0$ . In the following we use three different significance values  $\alpha = \{0.05, 0.01, 0.001\}$  to study the influence of this parameter on the obtained results.

#### B. Empirical adjustment of significance levels

The approach presented in the previous section does not rely on any kind of prior information from other sources, e.g., the literature, but uses only the information present in the data. In a machine learning terminology this is called an unsupervised method. The advantage of this method class namely no parameters need to be learned from training data is a disadvantage if the signal within the data is weak, because there is no direct way to beneficially influence the performance of the method by prior knowledge or information available in addition to the data set to be analyzed. Roughly speaking, for this reason, supervised methods are in general more powerful, because more information is used to analyze the data. As a first step extending the unsupervised method of the previous section to a supervised method we suggest to adjust empirically the significance level  $\alpha$  from the data and expert knowledge about the biological system under investigation from the literature. That means, we learn an efficient significance level  $\alpha_e$  indirectly by determining

$$t_{\alpha_e} = \begin{cases} t_\alpha - t_\alpha \cdot d_\alpha & : d_\alpha > 0 \\ t_\alpha & : else \end{cases} \quad (7)$$

from the data for selected gene interactions known to be present. Here  $d_\alpha$  is the mean value of the relative deviation

of  $t_\alpha$  from  $t_r$

$$d_\alpha = \left\langle \frac{t_\alpha - t_r}{t_\alpha} \right\rangle \quad (8)$$

The mean is evaluated with respect to some connections between genes that should lead to a pronounced (partial) correlation  $r$  and, hence, to  $t_r$  values larger than  $t_\alpha$ , because these interactions (edges in the network) are known from the literature. More precisely, we make the following assumptions. First, we assume that the data contain all information sufficient to infer all interactions between the genes and, hence, allow to reconstruct the gene networks, however, the signal is very weak and its detection difficult. Second, we assume that we know already reliably some gene interactions. For these genes Eq. 8 is estimated. If  $d_\alpha$  is less or equal to zero  $t_{\alpha_e}$  is not changed because all gene interactions can be detected from the data (high (partial) correlation coefficients)). If  $d_\alpha > 0$  the majority of these interactions would not be detected for a given significance level  $\alpha$ , because the null hypothesis would not be rejected. However, given our assumptions, this is wrong. The efficient significance level  $\alpha_e$  represents a correction of  $\alpha$  apparently not appropriate for the given data, because otherwise true (partial) correlation coefficients would be rejected. For our numerical simulations we use the genes given in table II chosen from the literature as 'true' gene interactions.

TABLE II

'TRUE' GENE INTERACTIONS ASSUMED TO BE PRESENT IN THE EXPERIMENTAL DATA.

SWI5 - CLB1
SWI5 - CLB2
CLB1 - CLN2
CLB6 - CLB2

#### IV. RESULTS

For our numerical studies we use the microarray data set of the *S.cerevisiae* cell cycle from SPELLMAN [14] and CHO [3]. The cell cultures were synchronized by three different methods, resulting in three different data sets alpha-factor, cdc15 and cdc28, and, hence, represent three different observations of the same biological process. We apply our method described in section III-A to all three data sets. Additionally, we apply the method to the combined data set.

Table III (top) shows the results for three different significance levels  $\alpha$  and Fig. 1 shows the corresponding values of  $d_\alpha$ . It is already clear from visual inspection that the center of mass of  $d_\alpha$  is for all three cases larger than zero. Numerically, we find  $d_{0.05} = 0.231$ ,  $d_{0.01} = 0.248$  and  $d_{0.001} = 0.191$ . This means, despite the fact that the gene interactions given in table II are necessary to ensure the progression of the yeast cell cycle they are undetectable given the corresponding significance levels  $\alpha$ . Again, we assume that this information taken from the literature is true and that the microarray data capture enough information about the underlying biological process that a detection should be possible.

TABLE III

ESTIMATED GENE INTERACTIONS FROM THE DATA OF THE YEAST CELL CYCLE. THE NUMBER OF VOTES INDICATES HOW MANY DATA SETS, ALPHA-FACTOR, CDC15, CDC28 OR THE COMBINED DATA SET, DETECT STATISTICALLY SIGNIFICANT AN INTERACTION BETWEEN THE EXPRESSION PROFILES OF GENE A AND B. TOP: SIGNIFICANCE LEVEL. BOTTOM: EFFICIENT SIGNIFICANCE LEVEL.

	$\alpha = 0.05$	$\alpha = 0.01$	$\alpha = 0.001$
votes 4	CLB6 - CDC53 CLB6 - HCT1 SIC1 - CDC34 SIC1 - SWI5 CDC53 - CDC20 CDC34 - SWI5	CLB6 - HCT1 SIC1 - SWI5 CDC53 - CDC20 CDC34 - SWI5	SIC1 - SWI5 CDC53 - CDC20 CDC34 - SWI5
votes 3	CLB6 - CDC20 SIC1 - CDC28 CDC28 - CDC34 CDC53 - HCT1 CDC20 - HCT1	CLB6 - CDC20 SIC1 - CDC34 CDC20 - HCT1	CLB6 - CDC20 CLB6 - HCT1 SIC1 - CDC34
votes 2	CLN1 - CLN3 CLB2 - CLB4 CLB2 - MCM1 CLB2 - SIC1 CLB2 - MBP1 CLB2 - CDC34 CLB6 - SIC1 CLB6 - CDC28 CLB6 - SKP1 CDC28 - SWI5 CDC28 - HCT1 CDC53 - SKP1 SKP1 - CDC20	CLN1 - CLN3 CLB2 - MCM1 CLB2 - MBP1 CLB2 - CDC34 CLB6 - SIC1 CLB6 - CDC28 CLB6 - CDC53 CLB6 - SKP1 SIC1 - CDC28 CDC28 - CDC34 CDC28 - HCT1 CDC53 - SKP1 SKP1 - CDC20	CLN1 - CLN3 CLB2 - MBP1 CLB2 - CDC34 CLB6 - SIC1 CLB6 - CDC53 CDC28 - HCT1 CDC53 - SKP1 SKP1 - CDC20
votes 4	CLB6 - HCT1 SIC1 - CDC34 SIC1 - SWI5 CDC53 - CDC20 CDC34 - SWI5	CLB6 - CDC53 CLB6 - HCT1 SIC1 - CDC34 SIC1 - SWI5 CDC53 - CDC20 CDC34 - SWI5	CLB6 - HCT1 SIC1 - SWI5 CDC53 - CDC20 CDC34 - SWI5
votes 3	CLB6 - CDC28 CLB6 - CDC53 CLB6 - CDC20 SIC1 - CDC28 CDC28 - CDC34 CDC20 - HCT1	CLB6 - CDC20 SIC1 - CDC28 CDC28 - CDC34 CDC53 - HCT1 CDC20 - HCT1	CLB6 - CDC20 SIC1 - CDC34
votes 2	CLN1 - CLN3 CLN3 - MBP1 CLB1 - CDC53 CLB2 - CLB4 CLB2 - MCM1 CLB2 - SIC1 CLB2 - MBP1 CLB2 - CDC34 CLB5 - SWI6 CLB5 - CDC20 CLB6 - SIC1 CDC28 - HCT1 CDC53 - SKP1 CDC53 - HCT1 SKP1 - CDC20	CLN1 - CLN3 CLB2 - CLB4 CLB2 - MCM1 CLB2 - SIC1 CLB2 - MBP1 CLB2 - CDC34 CLB6 - SIC1 CLB6 - CDC28 CDC28 - SWI5 CDC28 - HCT1 CDC53 - SKP1 SKP1 - CDC20	CLN1 - CLN3 CLB2 - MCM1 CLB2 - MBP1 CLB2 - CDC34 CLB6 - SIC1 CLB6 - CDC28 CLB6 - CDC53 SIC1 - CDC28 CDC28 - CDC34 CDC28 - HCT1 CDC53 - SKP1 SKP1 - CDC20

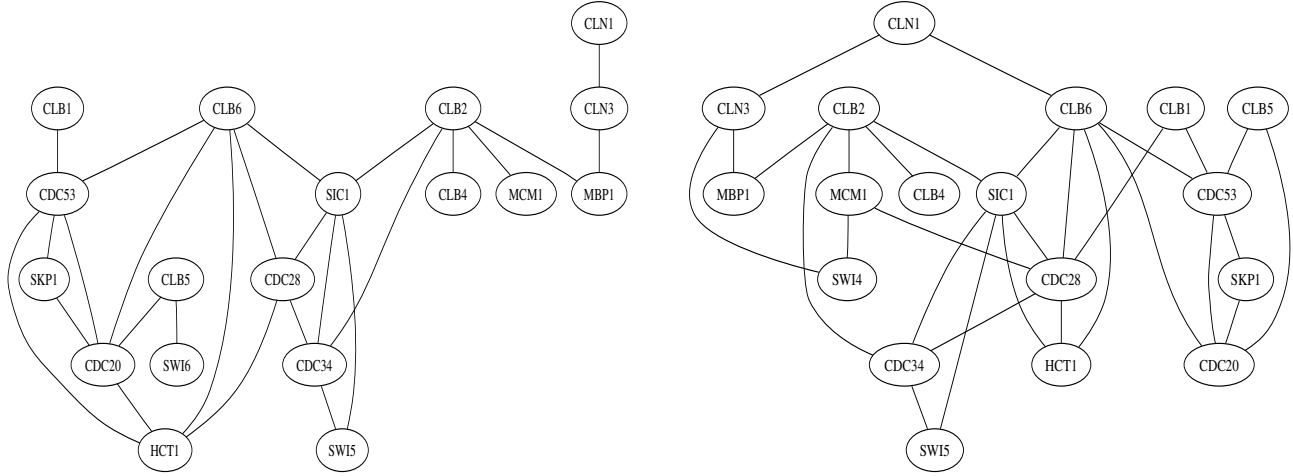


Fig. 2. Estimated networks structure of the gene interactions during the yeast cell cycle. Left: Visualization of the results from table III (bottom) for  $\alpha = 0.05$ . Right: Visualization for the combined data set for  $\alpha = 0.05$ .

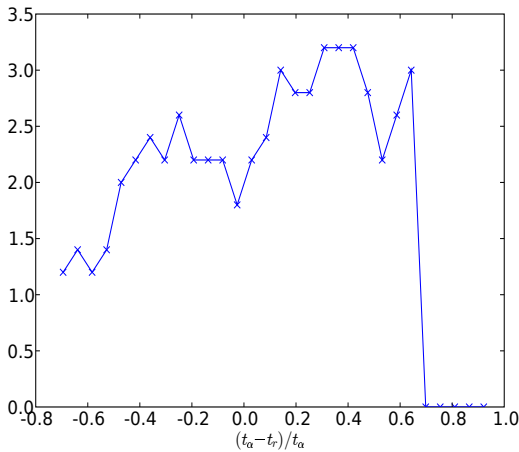
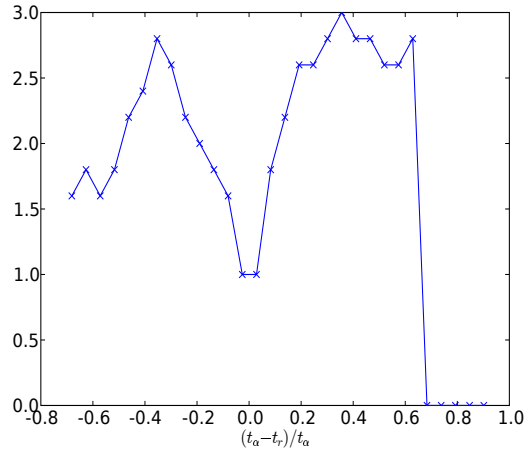
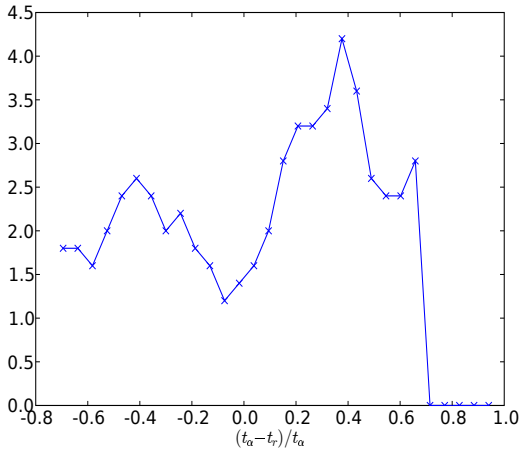


Fig. 1. Histograms of the (smoothed) relative deviation  $d_\alpha$  for  $\alpha = 0.05$  (top),  $\alpha = 0.01$  (middle) and  $\alpha = 0.001$  (bottom).

hence, in a larger number of gene interactions as expected. It is important to note that also for these cases the reconstructed networks are sparse, because only about 10% of the total number of possible gene-gene interactions pass the statistical test. This is remarkable, because we did not include any sparseness constraint in our algorithm. In Fig. 2 we show the estimated networks for  $d_{0.05}$  for the voted results in table III (bottom, left column) and the combined data set.

Finally, we study the influence of the maximal order  $O$  up to that we estimate the partial correlation coefficients. In Fig. 3 we show the number of edges in the network as function of  $O$ . The top figure corresponds to the uncorrected  $\alpha$  values and the bottom figure to  $\alpha_e$  for the results in table III. The number of edges changes most from  $O = 0$  to  $O = 1$ . This is reasonable, because  $O = 1$  is the first case that allows to exclude indirect causes between genes. However, it is surprising that, e.g., a further increase to  $O = 2$  seems to have almost no influence, suggesting that there are almost no parallel pathways genes

In table III (bottom) we present the results for the by Eq. 7 adjusted  $\alpha$  values. The overall observation is that the efficient significance value leads for all cases to a larger number of statistically significant (partial) correlation coefficients and,

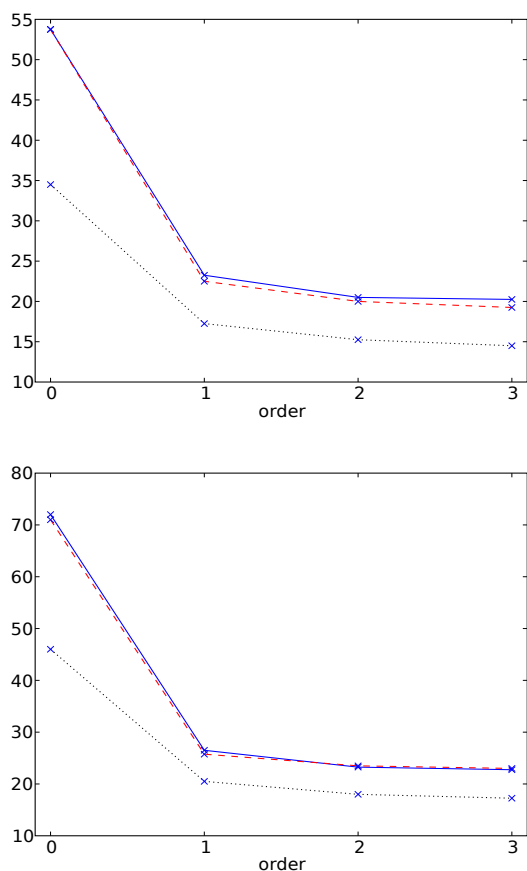


Fig. 3. Number of estimated edges present in the gene network in dependence on the order  $O$  of the partial correlation coefficients.

can interact through. This is an interesting point deserving more attention in further studies. Overall, these results clearly confirm [4] and justify the restriction of the maximal order up to partial correlation coefficients are evaluated.

## V. CONCLUSION

We introduced in this paper a supervised learning method to reconstruct a small subnetwork of gene interactions of the cell cycle of *S.cerevisiae* from microarray data. This method is similar to the PC [15] or IC algorithm [15], however, with two differences. First, for practical reasons we can only estimate the partial correlation coefficients up to a certain order. In this paper the maximal order was  $O = 3$ . We investigated the influence of the order on the number of estimated edges in the network and found that from  $O = 2$  to 3 the number of edges changes only slightly justifying a restriction of the maximal order to 2 or 3. This confirms previous results by [4]. Second, the PC as well as the IC algorithm are unsupervised methods which should work perfectly if all (partial) correlations present in the biological system can be estimated reliably. In reality, however, (partial) correlations present can not always be estimated from the data, e.g., because the data are too noisy. For this reason, we allow to learn the significance level from prior information not present in the data to at least partially

compensate such negative effects. This lead us to the definition of an efficient significance level  $\alpha_e$  which reduces the normal significance level  $\alpha$ . For example, for  $\alpha = 0.05$  we found from numerical simulations  $d_{0.05} = 0.231$ . This gives for the cdc15 data set  $t_{\alpha_e} = 1.595$  for  $df = N - 2 = 22$ . From a table of the Student's t distribution we find that  $t_{\alpha_e} = 1.595$  corresponds to a significance level  $\alpha_e$  of about 0.12. The crucial point is, normally one would not consider significance level above 0.05.

## ACKNOWLEDGMENT

We would like to thank Alberto de la Fuente and Bill Shipley for fruitful discussions.

## REFERENCES

- [1] H. Althoefer, A. Schleiffer, K. Wassmann, A Nordheim, and G. Ammerer. Mcm1 is required to coordinate g2-specific transcription in *saccharomyces cerevisiae*. *Mol Cell Biol.*, 15(11):5917–5928, 1995.
- [2] K. C. Chen, A. Csikasz-Nagy, B. Gyorfyy, J. Val, B. Novak, and J. J. Tyson. Kinetic analysis of a molecular model of the budding yeast cell cycle. *Mol Biol Cell*, 11:369–391, 2000.
- [3] R. Cho, M. Campbell, E. Winzeler, L. Steinmetz, A. Conway, L. Wodicka, T. Wolfsberg, A. Gabrielian, D. Landsman, D. Lockhart, and R. Davis. A genome-wide transcriptional analysis of the mitotic cell cycle. *Molecular Cell*, 2:65–73, 1998.
- [4] Alberto de la Fuente, Nan Bing, Ina Hoeschele, and Pedro Mendes. Discovery of meaningful associations in genomic data using partial correlation coefficients. *Bioinformatics*, 20(18):3565–3574, 2004.
- [5] Michael B. Eisen, Paul T. Spellman, Patrick O. Brown, and David Botstein. Cluster analysis and display of genome-wide expression patterns. *PNAS*, 95(25):14863–14868, 1998.
- [6] D. Geiger, T. S. Verma, and J. Pearl. Identifying independence in bayesian networks. *Networks*, 20:507–534, 1990.
- [7] M. G. Goebel, L. Goetsch, and B. Byers. The *ubc3* (*cdc34*) ubiquitin-conjugating enzyme is ubiquitinated and phosphorylated in vivo. *Mol Cell Biol*, 14:3022–3029, 1994.
- [8] M. Kanehisa and S. Goto. Kegg: kyoto encyclopa of genes and genomes. *Nuclei Acids Res.*, 28:27–30, 2000.
- [9] M. Koranda, A. Schleiffer, L. Endler, and G. Ammerer. Forkhead-like transcription factors recruit *ndd1* to the chromatin of g2/m-specific promoters. *Nature*, 406:94–98, 2000.
- [10] Judea Pearl. *Probabilistic Reasoning in Intelligent Systems*. Morgan-Kaufmann, 1988.
- [11] Judea Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge, 2000.
- [12] D. J. Sheskin. *Handbook of Parametric and Nonparametric Statistical Procedures*. RC Press, Boca Raton, FL, 3rd edition, 2004.
- [13] Bill Shipley. *Cause and Correlation in Biology*. Cambridge University Press, 2000.
- [14] P. Spellman and et al. Comprehensive identification of cell cycle-regulated genes of the yeast *saccharomyces cerevisiae* by microarray hybridization. *Mol. Biol. of the Cell*, 9:3273–3297, 1998.
- [15] P. Spirtes and C. Glymour. A algorithm for fast recovery of sparse causal graphs. *Social Science Computer Review*, 9:62–72, 1991.
- [16] P. Spirtes, T. Richardson, C. Meek, R. Scheines, and C. Glymour. Using path diagrams as a structural equation modelling tool. *Sociological Methods and Research*, 27:182–225, 1998.
- [17] Jeremy H. Toyn and et al. The *swi5* transcription factor of *saccharomyces cerevisiae* has a role in exit from mitosis through induction of the cdk-inhibitor *sic1* in telophase. *Genetics*, 145:85–96, 1996.
- [18] T Verma and J. Pearl. Causal networks: semantics and expressiveness. In *Proceedings of the 4th workshop on uncertainty in artificial intelligence*, pages 352–359. Mountain View CA, 1988.