國立臺灣大學生農學院生物產業機電工程學系

碩士論文

Department of Bio-Industrial Mechatronics Engineering

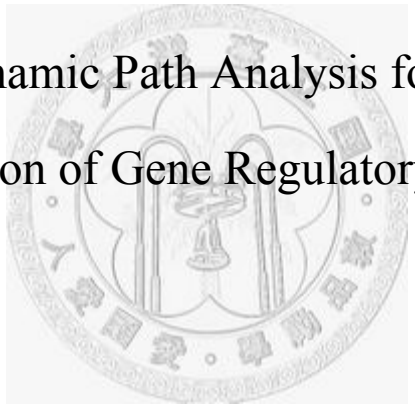
College of Bioresources and Agriculture

National Taiwan University

Master Thesis

利用動態路徑分析法進行基因調控網路鑑別與修復

Application of Dynamic Path Analysis for Identification and Modification of Gene Regulatory Networks



洪國智

Hung, Kuo-Chih

指導教授：周瑞仁 博士

Advisor: Chou, Jui-Jen, Ph.D.

中華民國97年6月

June, 2008

國立臺灣大學碩士學位論文
口試委員會審定書

利用動態路徑分析法進行基因調控網路鑑別與修復
Application of Dynamic Path Analysis for Identification and Modification of Gene Regulatory Networks

本論文係洪國智君（R94631012）在國立臺灣大學生物產業機電工程學系、所完成之碩士學位論文，於民國 97 年 06 月 16 日承下列考試委員審查通過及口試及格，特此證明

口試委員：　　（簽名）

（指導教授）

系主任、所長　　（簽名）

誌謝

感謝 周瑞仁教授在學生的碩士生涯中給予完整的指導，學生非常的感激。在您的嚴格指導下，學生學到了非常多做研究應該要有的態度及觀念。由於有著老師您的教導，使得學生在研究所的生涯中，過得非常充實。

另外，也要感謝 阮雪芬、黃宣誠、蕭介宗以及陳倩瑜教授撥空擔任學生的口試委員，並給予學生適當的指導及意見，使得本論文能夠更加的完善。

同時，也感謝 人毅以及振翔學長的特別照顧，使得學生在一開始能夠較快的熟悉實驗室的生活，以及在低潮的時候給予適當的鼓勵及建議，跟學長在一起的生活真的很快樂。也感謝 正宗學長的照顧及陪伴，與您打桌球及在實驗室嬉鬧的時光真得很快樂。也感謝 友俊及朝麟學長給予課業及研究上的意見，由於您們深厚的專業知識，使得學生見識到所謂的『專業』。另外，感謝同學 昌佑及俊翰一起陪伴度過了學生的研究生涯，一起睡在實驗室，一起買東西，一起出去玩，能夠認識你們真得很好。也感謝實驗室學弟 宗正、煥評以及遠霖，雖然你們對我不是很尊敬（呵呵），實驗室有著你們，變得熱鬧了許多，研究的氣氛也比較好也感謝我的家人時時刻刻不斷的給予關懷及幫助，沒有你們的幫忙，學生不會這麼順利的完成這個論文。

最後，感謝一路上曾經幫過我的大家，因為有著你們，一切將變得更好。謹以此文獻給大家。

摘要

本研究延伸路徑分析演算法(Path Analysis)，並引用微陣列晶片之動態資料，應用於資料庫中基因調控網路之鑑別上，以提供使用者判定基因調控網路可信度之參考。此外本研究亦利用演算法所提供的模型修正指標，適度地修正缺失的網路。

貝式模型、布林模型、結構方程模式與微分方程模型為現今較常見的基因調控網路模型。這些模型的目的係重建基因調控網路。本研究與上述方法不同之處在於著重於現有基因網路的鑑別。方法為根據 KEGG 的靜態網路產生多組候選的靜態網路，並拓展成 N 階動態網路。再計算這些動態網路之四組網路評估指標，以獲得最佳評估指標數目最多之網路，將之認定為最佳網路。

為了驗證本方法的可行性，本研究初步擷取十組基因調控網路進行鑑別。所採用的基因調控網路與微陣列晶片資料分別取自於 KEGG 與 NCBI 資料庫。網路為 cell cycle –yeast 之子網路，包括 regulation of autophagy 、MAPK signaling networks，資料為 Segal 等人所提供的微陣列晶片資料。此外，為了進一步的驗證本方法的可靠性，本研究也與 SSEM 演算法以及貝式建模相比較。

實驗結果顯示，十組測試基因調控網路中，共有七組基因調控網路被判定為最佳網路，鑑別率達 70%。另在針對該七組最佳網路，分別產生一些缺陷網路進行修正，其修正率為 43%。此外，本研究方法也與 SSEM 演算法及貝式建模比較，針對相同網路，本研究方法皆有較佳的網路鑑別性，對於各網路之有向性連結之判斷準確性至少達 60%以上。

關鍵字：路徑分析法、基因調控網路、網路評估指標，修正指標

Abstract

The study expands Path Analysis (PA), and adopts microarray time course data to identify gene regulatory networks (GRNs). It provides users degrees of confidence on GRNs in databases. In addition, defective networks can be modified based on modification indices in PA.

A couple of approaches, such as Bayesian, Boolean, structural equation modeling and differential equations model, are used for the reconstruction of gene regulatory networks in databases. We generate several alternative networks as candidates from original networks in KEGG database for comparison. Furthermore, the static networks are expanded to dynamic form (multiple orders). Finally, path analysis may suggest the best one in the network pool based on various performance indices. In other words, this approach can evaluate the existing networks in gene networks databases and provide users degrees of confidence on each network.

The gene regulatory networks are form KEGG in this study, including sub-networks of cell cycle –yeast, e.g., regulation of autophagy and MAPK signaling networks, and corresponding microarray time course data are adopted from NCBI database. Furthermore, we compare our approach with SSEM algorithm and dynamic Bayesian model method.

Seven out of ten original GRNs in KEGG are ranked the best networks by our approach, and 43 percent of defective networks generated from seven best original GRNs can be correctly modified. Besides, we obtain better results than SSEM algorithm and the dynamic Bayesian model method do to the same networks. The true positive rate on the directed links of networks is at least 60 percent.

Key Words: Path Analysis, gene regulatory networks, performance index, modification index.

Table of Contents

List of Figures

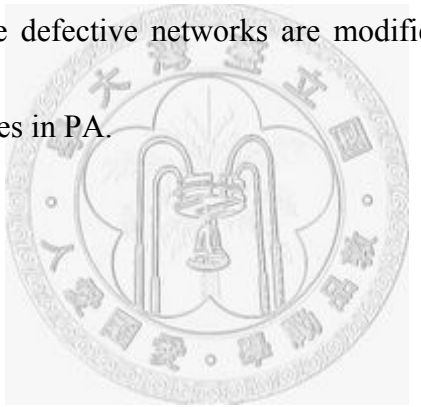
List of Tables

Chapter 1 Introduction

One of the most challenging works of systems biology is to reconstruct or identify structures and mechanisms of interaction among components of gene regulatory networks (GRNs) from microarray time course data. Because microarray time course data have dynamic information about gene expression, GRNs should be dynamic network to hold the temporal information about relationships among genes. Understanding GRNs can provide new notions for treating complex disease and breakthroughs for designing new drugs. Recently, several approaches have been proposed to reconstruct GRNs from microarray time course data, such as dynamic Bayesian networks (DBN), Boolean networks, structural equation modeling (SEM), neural network, and genetic algorithm. These approaches are used to search better match between reconstructed models from microarray time course data and GRNs in databases. The problem, mismatch between the models used for network inference and the real mechanisms used for data generation, may lead to various network structures or low accuracy of reconstructed model (Novikov and Barillot, 2008). Therefore, how to improve the accuracy of reconstructed model and to identify the reconstructed model is important issues.

Path analysis (PA), which helps understand comparative strengths of direct and indirect relationships among a set of variables, was first developed for ecology and later gradually adopted by social scientists, psychologists, and economists. Santos *et al*. (2005) applied PA to elucidate biochemical pathway, but the model of PA he used was static in essence.

In this study, we expanded path analysis (PA) to dynamic form and applied it to GRN identification. We can judge whether the original GRN is the proper network by four indices in PA, and the defective networks are modified toward the original one based on modification indices in PA.



Chapter 2 Literature Review

Biological networks usually represented by graph can be used to describe complex mechanism in biological systems, where nodes may indicate genes, proteins, metabolites or other-level modules and edges may describe relations among nodes. Many researchers focus on biological networks studies, particularly in the representation, reconstruction, identification or dynamic behaviors of the networks (Forster *et al.*, 2003; Herrgard *et al.*, 2006; Kim *et al.*, 2007; Gille *et al.*, 2007; Sotiropolos *et al.*, 2007). For the investigation of reconstructing GRNs based on gene expression data and dynamic form, Bayesian network (Murphy and Mian, 1999; Kim *et al.*, 2003; Van Berlo *et al.*, 2003; Zou *et al.*, 2005; Luna *et al.*, 2007), Boolean network (Mehra, 2004) and structural equation modeling (SEM) (Shieh *et al.*, 2008) were developed. Murphy and Mian (1999) first proposed a dynamic Bayesian network (DBN) to theoretically model various types of gene expression data. In this study, different strategies were proposed under various modeling environment such as partially observed parameters, fully observed parameters, partly known network structures and known structures. Kim *et al*. (2003) advanced a general framework of dynamic Bayesian networks to model real time course microarray data. They setup

criteria for network learning where biological knowledge was considered in posterior probabilities calculation. Luna *et al*. (2007) developed the variational Bayesian structure expectation maximization (VBSEM) algorithm that can learn posterior distribution of model parameters, and they also compared their results with KEGG. Boolean network is a simple computational model which simplifies large systems and provides insight into the overall behavior of biological networks. Mehra (2004) used Boolean approach to reconstruct the gene regulatory networks that can reduce the miss-predicted ratio by distinguishing direct casual relations and indirect casual relations in the networks. Shieh *et al*. (2008) alleged a stepwise structural equation modeling (SSEM) algorithm that concerned about latent variables in gene networks. Compared with other kinds of algorithms based on Bayesian networks or multivariate autoregressive model, SSEM obtains higher percentage accuracy on reconstructing networks.

For path analysis approach, Santos *et al*. (2005) first applied it to amplify the biochemical network of carotenoid, and their study showed the efficacy of path analysis for identifying key relation in complex pathways. Qiu *et al*. (2007) used path analysis to identify causal interrelationship of various biomarkers, and their results indicated that Coke-Oven exposure (cokework) played a more important role than the

genotypes in the induction of early genetic damage.

An S-system model (Savageau, 1969; Voit, 2000), a generalized framework for modeling and analyzing a biological system, is widely used to represent the mathematical relationship among components in the biological system (Wessels *et al*., 2001). In the study, it is employed to generate rich data for preliminary model identification under either noise-free or noise environment. If we know the interactions among genes described by S-system model, the corresponding data of genes can be produced.

SEM is a popular research tool in social science. It is a combination of two major statistical research fields: factor analysis and path analysis. In this study, we apply the SEM-approached path analysis (SEM-based PA) method to identify gene regulatory networks based on microarray time course data and dynamic model. Besides, defective networks can be modified toward identifiable GRNs based on modification indices in PA.

Chapter 3 Materials and Methods

The procedures of the network identification and defective network modification are shown in Figure 3-1. We first adopt KEGG static networks as our original networks, and generate several alternative networks as candidates for comparison. Based on a network structure from KEGG and time course data from NCBI, we further expand the static networks to dynamic ones, where the order of dynamic networks is equal to the number of time points in the dataset. Finally, path analysis may suggest the best one among those networks by four performance indices, and may modify defective networks by modification indices. In addition, we can add a gene to identified network, and generate extended networks. In the end, path analysis can judge the best one from extended networks.

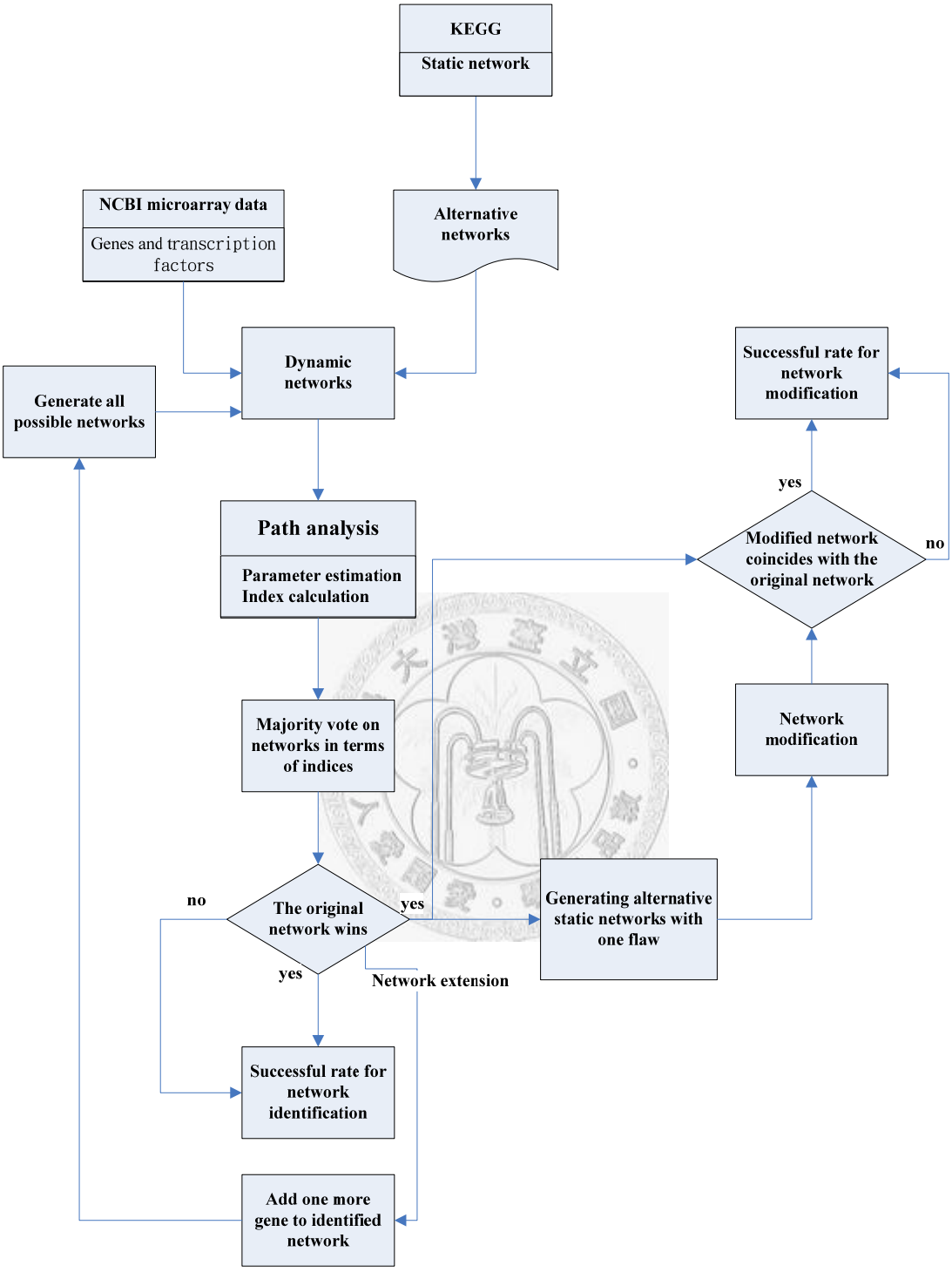


Figure 3-1 Flowchart of network identification and modification

3-1 Alternative networks

A series of alternative networks are generated as candidates to be compared with the original network. Let's take a simple example to illustrate the process of generation. The network shown in Figure 3-2 (a) is assumed as the designated KEGG static network with 4 genes and 3 links. The alternative networks are generated by connecting any 2 ones among 4 genes with 3 directional links and non-recursive loops. In this case, we obtain 19 alternative networks by Eq. (3-1), and those alternative networks are shown in Figure 3-2 (b). G, L, Q and T in Eq.(3-1) indicate the total numbers of genes, links, source genes and alternative networks correspondingly. In the example, there are 3 links in this network, and a 4 x 4 gene interaction matrix is build to indicate all relations among any 2 of 4 genes. The number of elements in the lower triangular of the matrix is $\frac{4 \bullet (4-1)}{2} = 6$. Q is equal to 1 in this case because only one gene1 (g_1) is assigned as source which is not affected by other genes in the network. Therefore, we obtain the number of alternative networks, T=C_3^6-1=19, by Eq. (3-1). All networks have their corresponding lower triangular matrices. Figure 3-3 shows the gene interaction matrices corresponding to their network configurations in Figure 3-2. In the Figure, g_1, g_2, g_3 , g_4 and L_{ij} represent gene1, gene2, gene3, gene4 and relation for g_j being affected by g_j.

$$T = C_{L}^{\frac{G(G-1)-Q(Q-1)}{2}} - 1 \qquad (3\text{-}1)$$

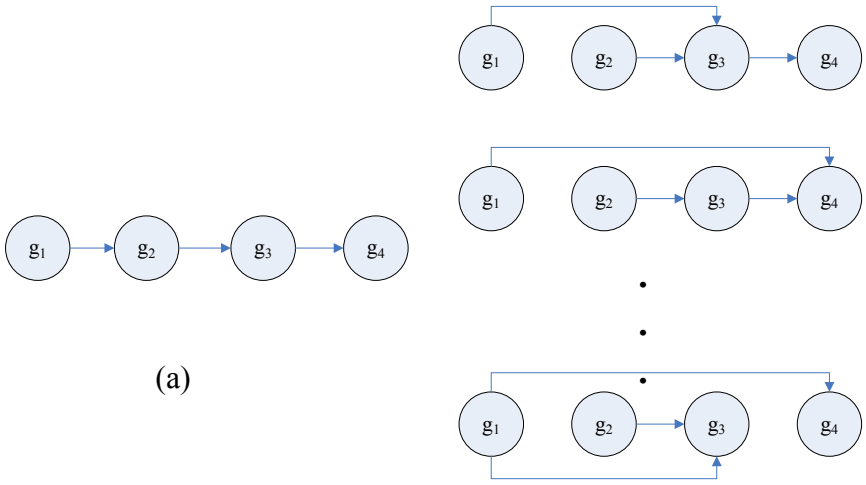


Figure 3-2 (a) Original network (b) Alternative networks

$$
\begin{array}{c} \\ g_1 \\ g_2 \\ g_3 \\ g_4 \end{array}
\begin{array}{c}
\begin{array}{cccc} g_1 & g_2 & g_3 & g_4 \end{array} \\
\begin{bmatrix} 0 & 0 & 0 & 0 \\ L_{12} & 0 & 0 & 0 \\ 0 & L_{23} & 0 & 0 \\ 0 & 0 & L_{34} & 0 \end{bmatrix}
\end{array}
$$

(a)

$$
\begin{array}{c} \\ g_1 \\ g_2 \\ g_3 \\ g_4 \end{array}
\begin{array}{c}
\begin{array}{cccc} g_1 & g_2 & g_3 & g_4 \end{array} \\
\begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ L_{13} & L_{23} & 0 & 0 \\ 0 & 0 & L_{34} & 0 \end{bmatrix}
\end{array}
\qquad
\begin{array}{c} \\ g_1 \\ g_2 \\ g_3 \\ g_4 \end{array}
\begin{array}{c}
\begin{array}{cccc} g_1 & g_2 & g_3 & g_4 \end{array} \\
\begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & L_{23} & 0 & 0 \\ L_{14} & 0 & L_{34} & 0 \end{bmatrix}
\end{array}
$$

$$
\vdots
$$

$$
\begin{array}{c} \\ g_1 \\ g_2 \\ g_3 \\ g_4 \end{array}
\begin{array}{c}
\begin{array}{cccc} g_1 & g_2 & g_3 & g_4 \end{array} \\
\begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ L_{13} & L_{23} & 0 & 0 \\ L_{14} & 0 & 0 & 0 \end{bmatrix}
\end{array}
$$

(b)

Figure 3-3 Gene interaction matrices in Figure 3-2

3-2 Dynamic networks expanded from static networks

The original and alternative static networks are further expanded to an n-order dynamic form, where n indicates the number of time course. In other words, if gene y_i affects gene y_j in a static network, gene y_i in time t may affect gene y_j at time point $t, t+1, \cdots,$ or/and $t+(n-1)$. Figure 3-4 shows a dynamic network sample. In this network, gene expression level $y_i(t-1)$ may affect $y_j(t-1)$, $y_j(t)$ or/and $y_j(t+1)$. In the same token, $u_k(t)$, the transcription factor, may regulate other genes at different time points.

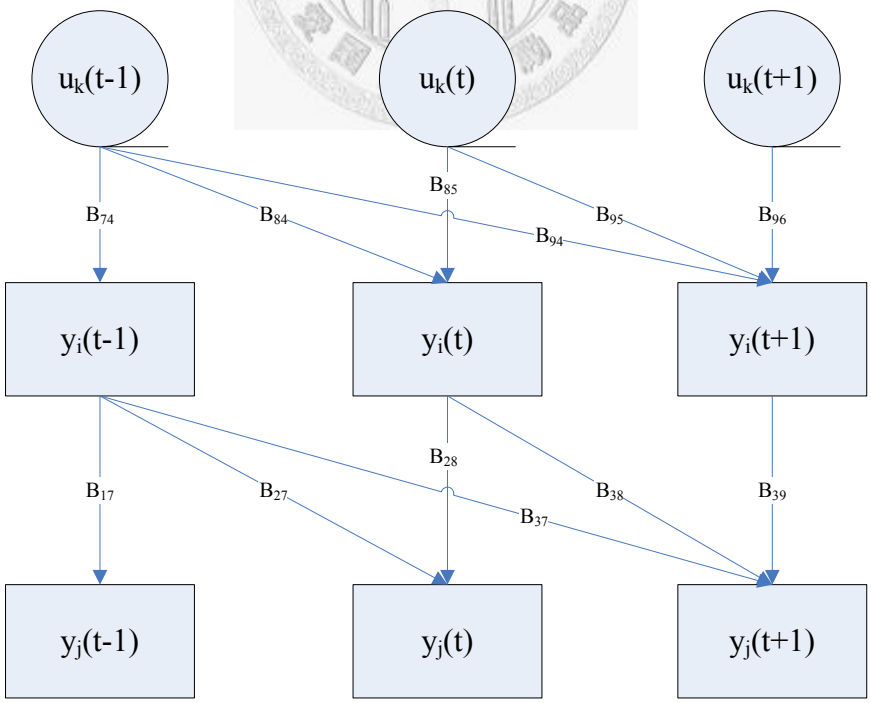


Figure 3-4 Dynamic network

3-3 Data interpolated

In microarray time course data, only one data is provided at each time point. Therefore, data augmentation is necessary for network identification. In this study, third-order curve fitting (piecewise cubic spline) is applied to generate extra data needed for modeling. 9 adjacent points around the gene expression value at each time point are obtained from piecewise cubic spline. Figure 3-5 shows the microarray data at time 0, 5, 15, 30 and 60 with corresponding values -0.464, -1.168, 0.162, -0.399 -0.89.

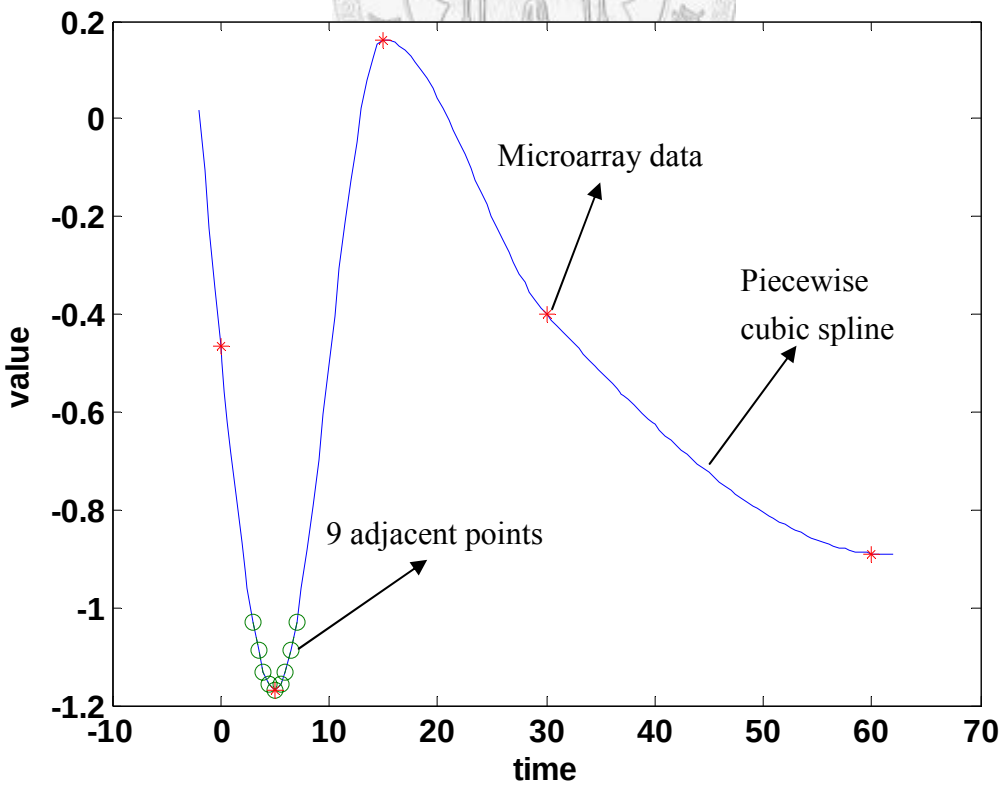


Figure 3-5 Interpolated data

3-4 Path analysis algorithm and parameter estimation

The measurement equation (Eq. (3-2)) plays a major role in describing a gene network for path analysis (SEM approach).

Measurement equation: $Y = WY + E$ (3-2)

Where

$$Y = \begin{bmatrix} TF_m(t) \\ ExoG_n(t) \\ EndoG_r(t) \end{bmatrix}$$

$TF_m(t)$: A vector of transcription regulators, $1 \leq m \leq M$.

$ExoG_n(t)$: A vector of exogenous genes, $1 \leq n \leq N$.

$EndoG_r(t)$: A vector of endogenous genes, $1 \leq r \leq R$.

and

M: Number of transcription regulators.

N: Number of exogenous variables (genes).

R: Number of endogenous variables (genes).

E: Residual terms of all observed variables.

W: Weight matrix of all observed variables.

In the dynamic gene network, we consider genes and corresponding transcription

factors of gene's level as their variables and links representing their relations. All variables in the network model are observable. We derive a model implied covariance matrix (Eq. (3-3)) from the measurement equation, where unknown parameter vector θ consists of weighting matrix, W, and the covariance matrix of E, Ψ.

$$\begin{aligned} Y &= WY + E \\ &= (I - W)^{-1} E \end{aligned}$$

Assume that Σ and Ψ are the covariance matrices of Y and E correspondingly.

$$\begin{aligned} \mathrm{Cov}(Y) \equiv \Sigma(W, \Psi) &= \Sigma(\theta) \\ &= (I - W)^{-1} E[(I - W)^{-1} E]^T \\ &= (I - W)^{-1} EE^T (I - W)^{-T} \\ &= (I - W)^{-1} \Psi (I - W)^{-T} \end{aligned} \qquad (3\text{-}3)$$

The Maximum likelihood (ML) algorithm is applied to estimate parameter vector θ by minimizing the discrepancy function shown in Eq. (3-4). Given sample covariance matrix S which is a Wishart distribution, we have the probability density function of S.

$$f(S|\Sigma_0) = \frac{C}{|\Sigma_0|^{\frac{R-1}{2}}} \exp\{\frac{-(R-1)}{2} \mathrm{tr} S \Sigma_0^{-1}\}$$

Where $\Sigma_0 = \Sigma(\theta_0) : \theta_0$ is a true parameter vector but unknown

C : normalized constant

The negative log-likelihood function is as follows:

$$-\log C + \frac{(R-1)}{2}[\log|\Sigma(\theta)| + \mathrm{tr} S \Sigma(\theta)^{-1}]$$

Hence the discrepancy function is given as follows (Jöreskog, 1978):

$$F(\theta) = \log|\Sigma(\theta)| + trS\Sigma(\theta)^{-1} - \log|S| - Z$$

Where Z is equal to M + N + R, number of all observable variables.

Because we also consider the minimization between sample mean vector $\bar{y}$ and the model implied mean vector $\mu(\theta)$, the discrepancy function is now become

$$F_m(\theta) = \log|\Sigma(\theta)| + trS\Sigma(\theta)^{-1} - \log|S| - Z + (\bar{y} - \mu(\theta))^T \Sigma^{-1}(\bar{y} - \mu(\theta)) \qquad (3\text{-}4)$$

To estimate θ, the gradient vector $dF_m(\theta)/d\theta$ and the Hessian matrix $d^2F_m(\theta)/d\theta^2$ are given to minimize $F_m(\theta)$. The measurement equation of Figure 3-4 can be described by Eq. (3-5).

$$\begin{bmatrix} y_j(t-1) \\ y_j(t) \\ y_j(t+1) \\ u_k(t-1) \\ u_k(t) \\ u_k(t+1) \\ y_i(t-1) \\ y_i(t) \\ y_i(t+1) \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & B_{17} & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & B_{27} & B_{28} & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & B_{37} & B_{38} & B_{39} \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & B_{74} & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & B_{84} & B_{85} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & B_{94} & B_{95} & B_{96} & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} y_j(t-1) \\ y_j(t) \\ y_j(t+1) \\ u_k(t-1) \\ u_k(t) \\ u_k(t+1) \\ y_i(t-1) \\ y_i(t) \\ y_i(t+1) \end{bmatrix} + E \qquad (3\text{-}5)$$

Where $y_j(t)$: the expression level of gene j at time t.

$u_k(t)$: the expression level of transcription factor k at time t.

E: 9×1 vector

3-5 Performance indices

In this study, four indices are considered for network identification, including weighted least square Chi-square (WLS χ^2) (Eq. (3-6)), Akaike information criterion (AIC) (Eq. (3-7)), standard root mean square residual (S-RMR) (Eq. (3-8)), and goodness-of-fit index (GFI) (Eq. (3-9)). For the best network, WLS- Chi-Square, AIC, and S-RMR indices are supposed to be small, and GFI index is supposed to be large.

$$\text{WLS}\,\chi^2 = c = (O-1)F_m(\theta) \tag{3-6}$$

$$\text{AIC} = c + 2p \tag{3-7}$$

$$\begin{aligned} S-\text{RMR} &= \text{Normalization (RMR)} \\ &= \text{Normalization}\left(\sqrt{2\sum_{i=1}^{M+N+R}\sum_{j=1}^{i}\frac{(S_{ij}-\sum(\theta)_{ij})^2}{(M+N+R)(M+N+R+1)}}\right) \end{aligned} \tag{3-8}$$

$$\text{GFI} = \frac{\text{tr}(\sum(\theta)^t W \sum(\theta))}{\text{tr}(S^t W S)} \tag{3-9}$$

Where O: sample size

p: the number of independent parameters estimated

M+N+R: the number of observed variables

3-6 Modification of networks

Modification index (MI) offered by LISREL 8.8 (Jöreskog and Sörbom, 2006) helps us further study the ability of correction on defective networks. In the illustrated example shown in Figure 3-6, we consider the network with only one missing link, which is between gene ATG7 and other genes. With the static defective network, we expand the static network to a dynamic one. In the study, we then focus on the defective part to obtain the suggestion on the potential links in the flawed part of the network by modification indices. The potential links between gene ATG7 and gene ATG10, where MI>9, are shown by solid lines in the dynamic defective network. Besides ATG10, We also examine the indices between gene ATG7 and other genes. Here, links with MI>9 are shown by dotted lines (e.g., gene ATG5 and ATG16). A link between two genes is added in the static defective network if the summation of MI values in the dynamic defective network is greater than a specified threshold. To calculate MI's more efficiently, we only adopt three links between two genes with top weights in the same layer of the defective part.

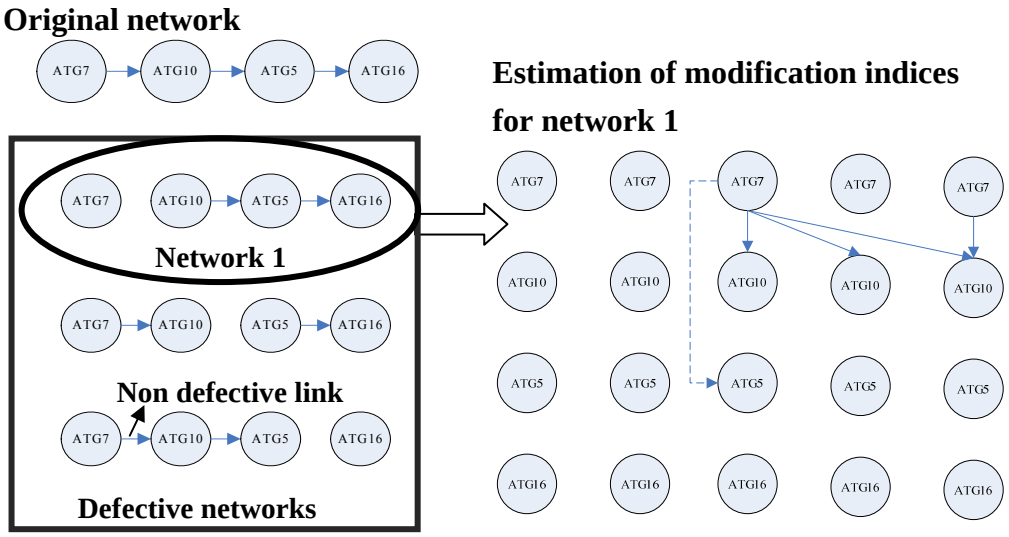


Figure 3-6 Determination of possible links in a defective network

3-7 Network extension

Network which has been identified can be extended and further identified by considering more genes possibly involving in the network. We first add one more gene to the network and generate all possible links for the new gene. Finally, path analysis in LISREL8.8 is applied and suggests the best one for the extended networks based on various performance indices. The number of extended networks can be calculated by Eq. (3-10), where NG is the total number of genes in the network prior to extending, and H is the total number of extended networks. For example, if we have one network, which has five genes, we can obtain nine extended networks by Eq. (3-10).

$$H = 2 \times NG - Q \tag{3-10}$$

Figure 3-7 shows a new gene, Clb1 and all possible links (A, B, C, D, E, F, G, H, and I). Solid lines represent given links in the network, and dotted lines indicate all feasible links. Figure 3-8 shows all the extended networks, and we can find the best network by path analysis.

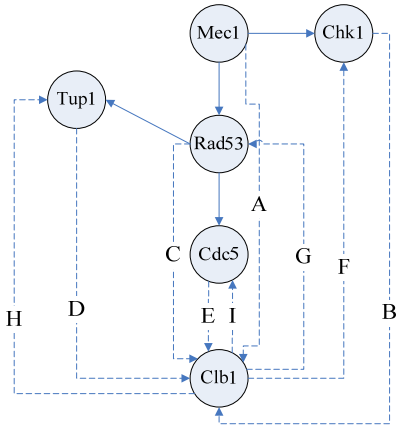


Figure 3-7 All possible links for the new gene, Clb1

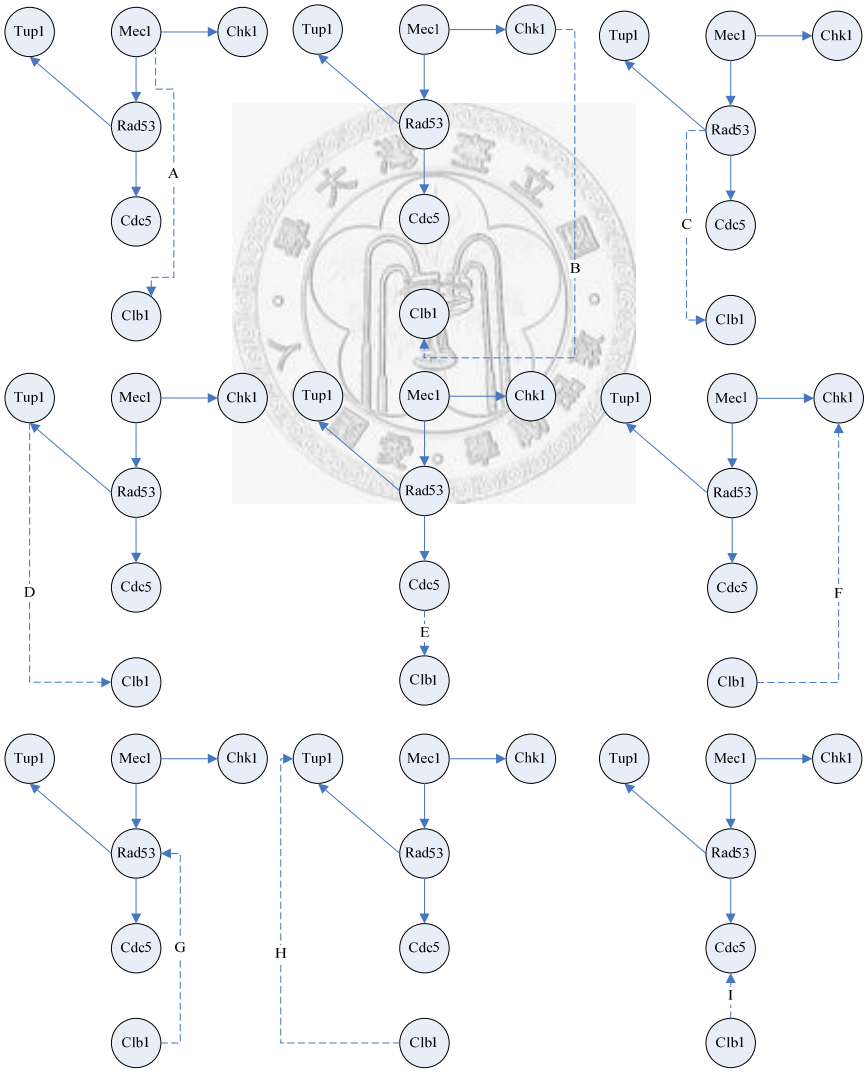


Figure 3-8 Extended networks

3-8 Data generation for simulation

For evaluating the effectiveness of PA applied in the identification of networks, we use S-system to describe network structures and generate time series data for simulation. The S-system model, a set of power-law equations, is shown in Eq. (3-11).

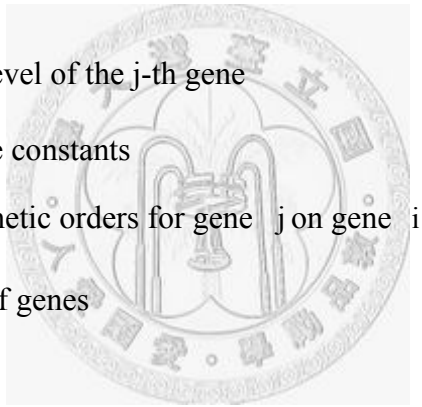
$$\dot{X}_i = \alpha_i \prod_{j=1}^{Z} X_j^{g_{ij}} - \beta_i \prod_{j=1}^{Z} X_j^{h_{ij}} \quad \text{for} \quad i = 1,2,...,n \tag{3-11}$$

in which: X_j : expression level of the j-th gene

α_i and β_i : rate constants

g_{ij} and h_{ij} : kinetic orders for gene j on gene i .

Z: the number of genes



We sample data every 0.5 time unit from S-system model. Two cases, shown in Figure 3-9, are designed to illustrate the identification of the networks with PA algorithm. A chain type network with 3 genes (case 1) and a branched type network with 3 genes (case 2) are considered.

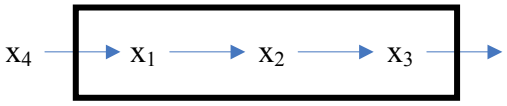
Case1

$$\dot{X}_1 = 0.5X_4 - 0.5X_1^{0.5} \qquad X_1(0) = 1$$

$$\dot{X}_2 = 0.5X_1^{0.5} - 4X_2 \qquad X_2(0) = 1$$

$$\dot{X}_3 = 4X_2 - 2X_3^{0.75} \qquad X_3(0) = 1$$

$$X_4 = 0.5$$



Case2

$$\dot{X}_2 = 5X_1^{0.5} - 4X_2^{0.5} \qquad X_1(0) = 0.1$$

$$\dot{X}_3 = 2X_2^{0.5} - 1.25X_3^{0.5} \qquad X_2(0) = 0.5$$

$$\dot{X}_4 = 8X_2^{0.5} - 3X_4^{0.5} \qquad X_3(0) = 0.9$$

$$X_4(0) = 0.75$$

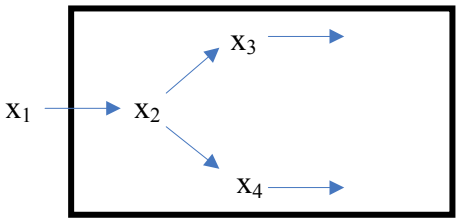


Figure 3-9 S-system model and networks of Case 1 and Case 2

Chapter 4 Results and Discussion

4-1 Networks and data from S-system

Tables 4-1 and 4-2 show the identification results of case 1 and case 2, and four indices such as WLS χ^2, AIC, S-RMR and GFI are evaluated. In these two cases, the original networks obtain majority votes (Figures 4-1 and 4-2). Therefore, they are considered the best ones among all candidates of gene network pool in terms of performance indices.

Table 4-1 Results of identification for case 1

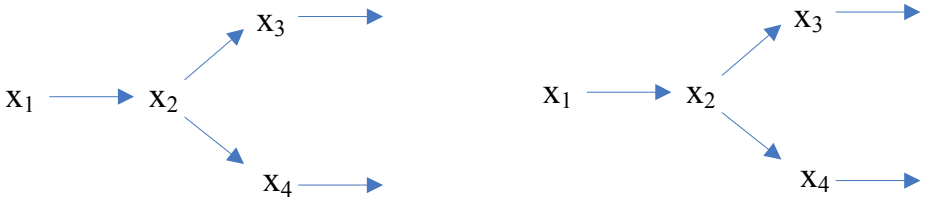
	Number of alternative networks	Number of performance indices with best scores	Results
Case 1	2	3/4	Match

Table 4-2 Results of identification for case 2

	Number of alternative networks	Number of performance indices with best scores	Results
Case 2	2	3/4	Match

(a) (b)

Figure 4-1 Case 1 (a) Original network; (b) Network with majority votes

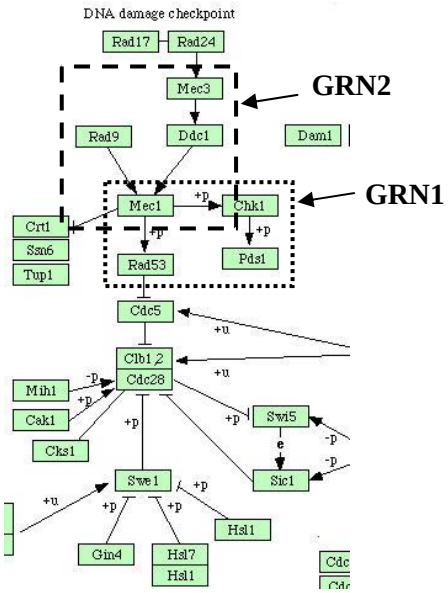


(a) (b)

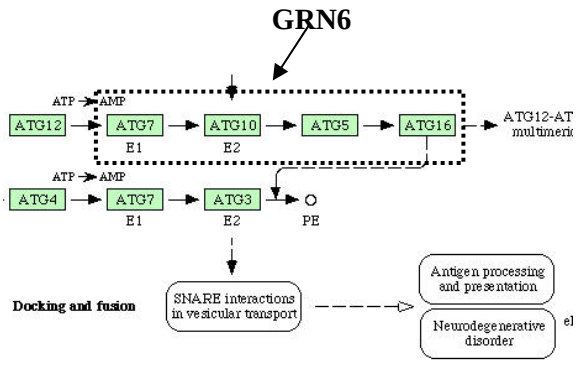
Figure 4-2 Case 2 (a) Original network; (b) Network with majority votes

4-2 Identification of networks in KEGG

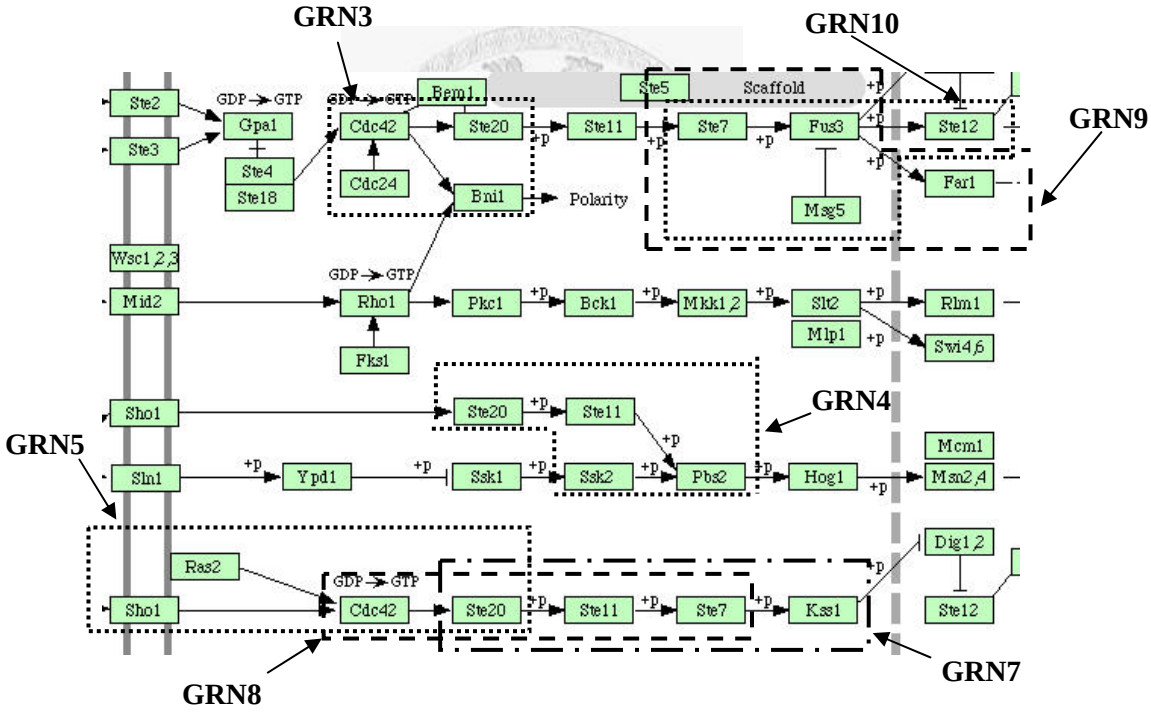
We further apply network identification approach to the networks in KEGG and data acquired from NCBI. Figure 4-3 shows ten sub-networks regarding yeast cell cycle may, including regulation of autophagy and MAPK signaling networks. Some networks may overlap, such as GRN1 and GRN2; GRN5, GRN7 and GRN8; GRN9 and GRN10. Network GRN1 includes four genes, Mec1, Chk1, Rad53 and Pds1. We express it as a set GRN1={Mec1, Chk1, Rad53 and Pds1}; GRN2 = {Mec3, Rad9, Ddc1, Mec1}; GRN 5={Ras2, Sho1, Cdc42, Ste20}; GRN7={Ste20, Ste11, Ste7, Kss1}; GRN8= {Cdc42, Ste20, Ste11, Ste7}; GRN9={Ste7, Msg5, Fus3, Far1}; GRN10 ={Ste7, Msg5, Fus3, Ste12}. Database YEASTRACT (Teixeira *et al*, 2006) offers the necessary information of transcription factors about their linked genes. Data with 5 time points available in NCBI are provided by Segal *et al*. (2003).



Cell cycle- Saccharomyces cerevisiae



Regulation of autophagy-Saccharomyces cerevisiae



MAPK signaling pathway-Saccharomyces cerevisiae

Figure 4-3 GRNs in KEGG

4-2-1 Identification results of ten networks

Table 4-3 shows the genes and its transcription factors, and blank space in an entry indicates that there exist no TFs in YEASTRACT. Table 4-4 shows number of alternative networks, number of performance indices with best scores and result in GRN1 ~ 10. The GRN1, GRN6, GRN7 and GRN8 have 19 alternative networks and GRN2, GRN3, GRN4, GRN5, GRN9 and GRN10 have 9 alternative networks. Figures 4-4 ~ 4-13 show the original network and the network, which wins the majority votes of performance indices in the network pool, respectively, in each Figure.

Table 4-3 Genes and their transcription factors for GRN1 ~ 10

Networks	Genes				Transcription factors			
GRN1	Mec1	Chk1	Rad53	Pds1	Yap1p		Arr1p	Fkh2p
GRN2	Mec3	Rad9	Ddc1	Mec1	Hap4p	Hac1p		Yap1p
GRN3	Cdc24	Cdc42	Ste20	Bnil	Abf1p	Upc2p	Ino4p	Hcm1p
GRN4	Ste20	Ssk2	Ste11	Pbs2	Ino4p	Leu3p	Gcn4p	
GRN5	Ras2	Sho1	Cdc42	Ste20	Ace2p	Mbp1p	Upc2p	Sok2p
GRN6	ATG7	ATG10	ATG5	ATG16	Aft1p	Aft1p	Aft1p	Put3p
GRN7	Ste20	Ste11	Ste7	Kss1	Sok2p	Gcn4p	Abf1p	Yap1p
GRN8	Cdc42	Ste20	Ste11	Ste7	Upc2p	Sok2p	Gcn4p	Abf1p
GRN9	Ste7	Msg5	Fus3	Far1	Abf1p	Mcm1p	Hap4p	Mcm1p
GRN10	Ste7	Msg5	Fus3	Ste12	Abf1p	Mcm1p	Hap4p	Dot6p

Table 4-4 Identification results for GRN1 ~ 10

Networks	Number of alternative networks	Number of performance indices with best scores	Results
GRN1	19	0/4	Mismatch
GRN2	9	4/4	Match
GRN3	9	0/4	Mismatch
GRN4	9	0/4	Mismatch
GRN5	9	3/4	Match
GRN6	19	4/4	Match
GRN7	19	4/4	Match
GRN8	19	4/4	Match
GRN9	9	3/4	Match
GRN10	9	4/4	Match

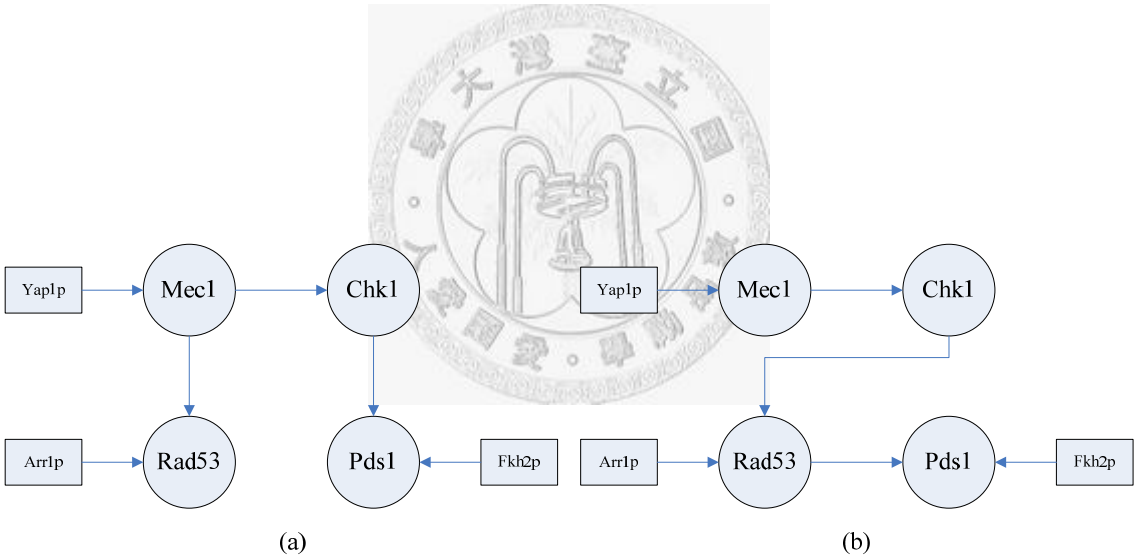


Figure 4-4 GRN1: (a) Original network; (b) Network with majority votes

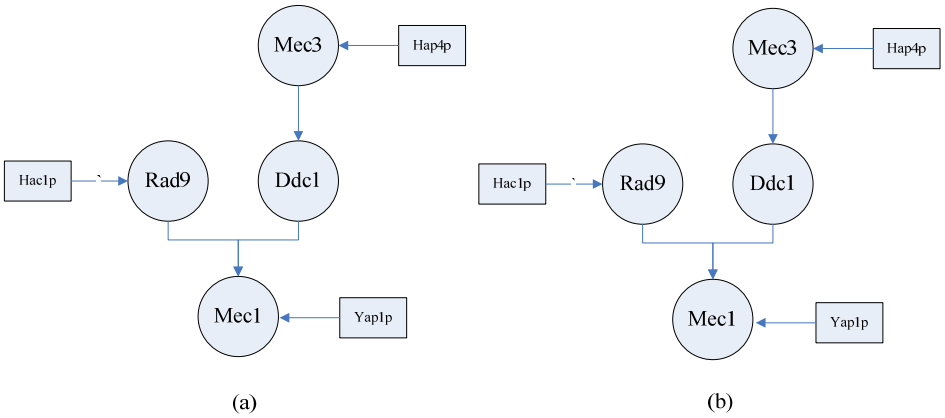


Figure 4-5 GRN2: (a) Original network; (b) Network with majority votes

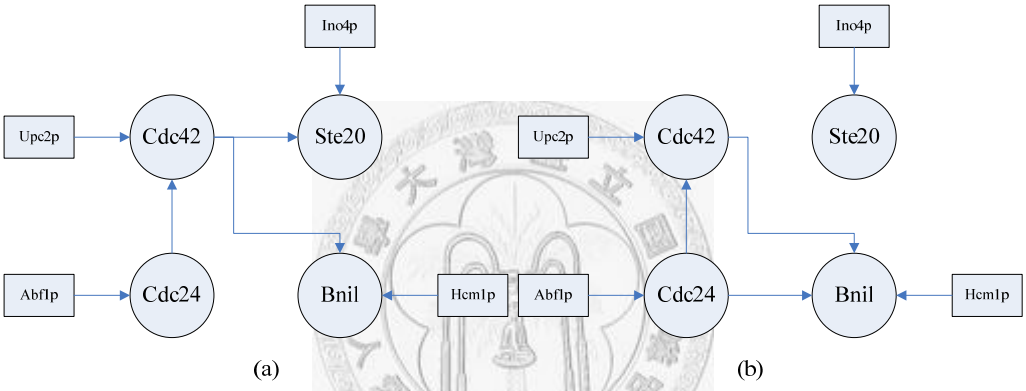


Figure 4-6 GRN3: (a) Original network; (b) Network with majority votes

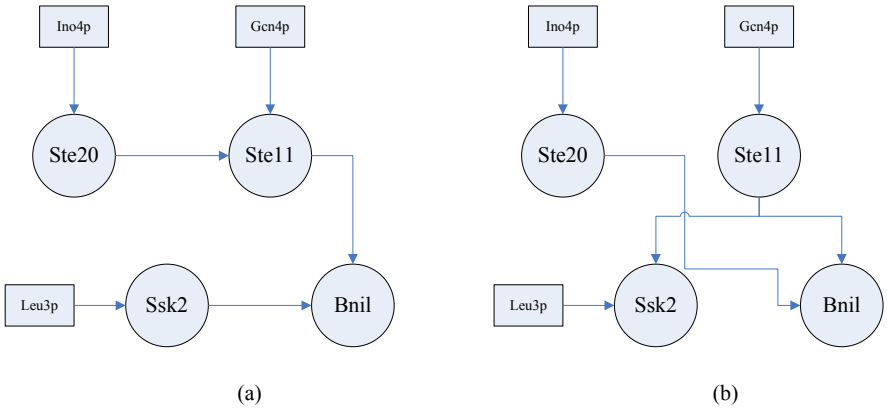
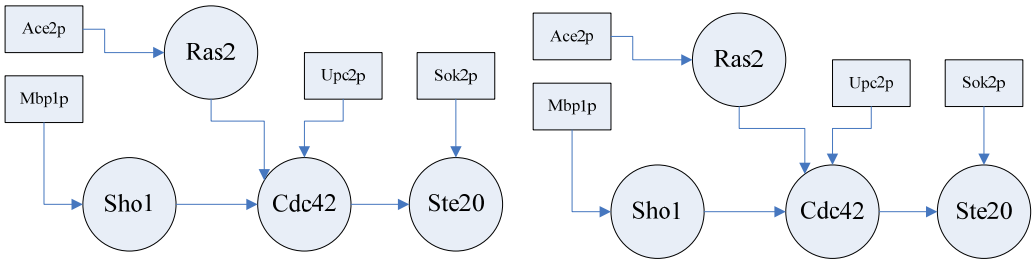
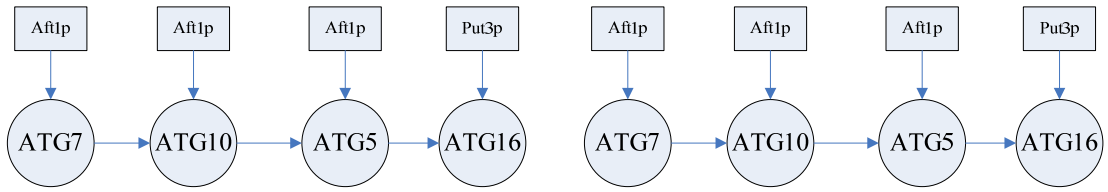


Figure 4-7 GRN4: (a) Original network; (b) Network with majority votes



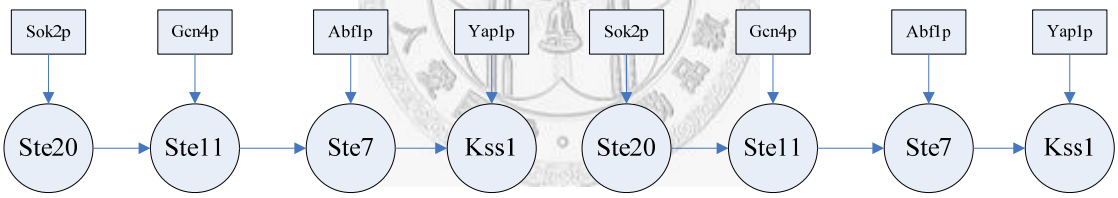
(a) (b)

Figure 4-8 GRN5: (a) Original network; (b) Network with majority votes



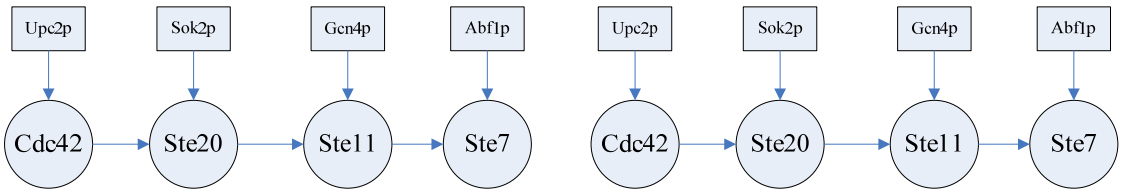
(a) (b)

Figure 4-9 GRN6: (a) Original network; (b) Network with majority votes



(a) (b)

Figure 4-10 GRN7: (a) Original network; (b) Network with majority votes



(a) (b)

Figure 4-11 GRN8: (a) Original network; (b) Network with majority votes

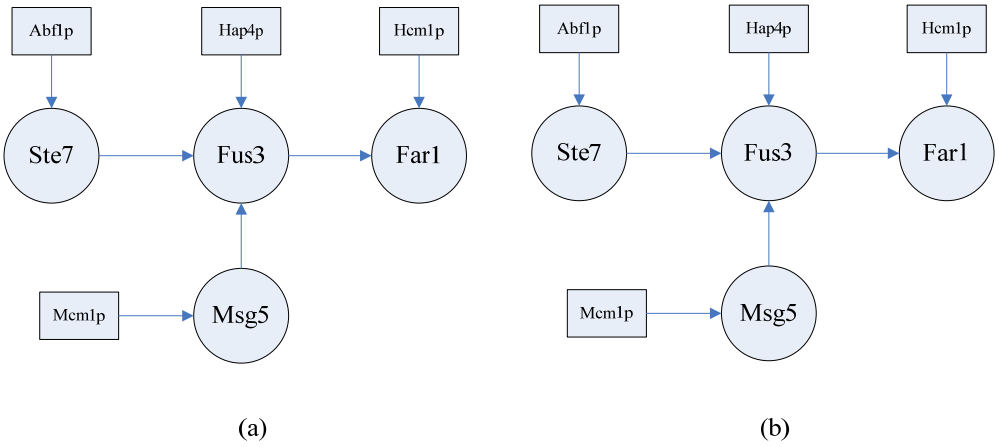


Figure 4-12 GRN9: (a) Original network; (b) Network with majority votes

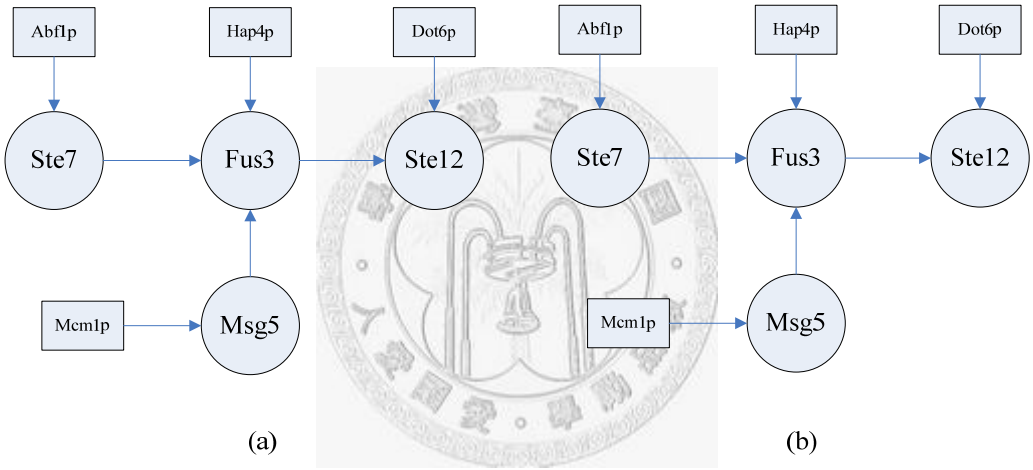


Figure 4-13 GRN10: (a) Original network; (b) Network with majority votes

Table 4-5 shows the results of identification, where 7 of 10 original networks wins the majority votes of performance indices in the network pool, including GRN2, GRN5, GRN6, GRN7, GRN8, GRN9 and GRN10. In the example of GRN10 shown in Table 4-6, the original network gets 4 best performance indices, where symbol √ indicates the best one in the specific index. Table 4-7 provides detailed index values for the 7 original

networks which are correctly identified, where symbol X indicates that the network doesn't obtain the best score among the network pool. Therefore, GRN5 and GRN9 don't perform better than alternative networks do in index S-RMR. The total numbers of best indices for each network are given in the bottom row of Table 4-7.

Table 4-5 Results of identification

Networks	Best one
GRN1	No
GRN2	Yes
GRN3	No
GRN4	No
GRN5	Yes
GRN6	Yes
GRN7	Yes
GRN8	Yes
GRN9	Yes
GRN10	Yes
Accuracy	70%

Table 4-6 GRN10: Networks marked with check symbols

GRN10	Original	AL1	AL2	AT3	AL4	AL5	AL6	AL7	AL8	AT9
WLS χ^2	√									
AIC	√									
S-RMR	√		√		√	√		√		
GFI	√		√		√	√		√		
Sum	4	0	2	0	2	2	0	2	0	0

"AL": Alternative network

Table 4-7 Index values for the 7 original networks

Index	GRN2	GRN5	GRN6	GRN7	GRN8	GRN9	GRN10
WLS χ^2	43.23	21.63	3.1	6.83	7.95	33.11	31.01
AIC	363.23	391.63	373.1	376.83	377.95	403.11	401.01
S-RMR	0.015	0.034(X)	0.049	0.0017	0.005	0.0081(X)	0.013
GFI	0.77	0.88	0.98	0.96	0.95	0.83	0.84
Votes	4/4	3/4	4/4	4/4	4/4	3/4	4/4

Furthermore, different models are compared in Table 4-8, including static model, static model with transcription factors, dynamic model, and dynamic model with transcription factors, where "Yes" indicate that the original network is the best one, and "No" indicate that the original network is not the best. From Table 4-8, the dynamic model with transcription factors is usually better than others.

Table 4-8 Comparison among different models

	Static	Static+TF	Dynamic	Dynamic+TF
GRN1	No	No	No	No
GRN2	No	No	No	Yes
GRN3	No	No	No	No
GRN4	No	No	Yes	No
GRN5	No	No	No	Yes
GRN6	No	No	No	Yes
GRN7	No	No	No	Yes
GRN8	No	No	No	Yes
GRN9	No	No	No	Yes
GRN10	Yes	No	No	Yes
Accuracy	10%	0%	10%	70%

"TF": Transcription factors

4-2-2 Results of modification

Table 4-9 shows the results of network modification. We create 21 defective networks in total from the original networks which have been correctly identified, and each defective network is allowed to have one flaw link. Overall, 9 of 21 defective networks can eventually be corrected.

Table 4-9 Results of modification

	Description	Suggestion	Results
GRN2	Flaw: Mec3 → Ddc1	Mec3 → Ddc1	Succeed in correction
	Flaw: Rad9 → Mec1	Rad9 → Mec1	Succeed in correction
	Flaw: Ddc1 → Mec1	Ddc1 → Mec1	Succeed in correction
GRN5	Flaw: Ras2 → Cdc42	Ras2 → Ste20	Fail to correct
	Flaw: Sho1→ Cdc42	Sho1→ Ste20	Fail to correct
	Flaw: Cdc42→ Ste20	Ras2 → Ste20	Fail to correct
GRN6	Flaw: ATG7 → ATG10	ATG7 → ATG10	Succeed in correction
	Flaw: ATG10 → ATG5	ATG10 → ATG16	Fail to correct
	Flaw: ATG5 → ATG16	ATG7 → ATG16	Fail to correct
GRN7	Flaw: Ste20 → Ste11	Ste20 → Ste11	Succeed in correction
	Flaw: Ste11→ Ste7	Ste20 → Ste7	Fail to correct
	Flaw: Ste7→ Kss1	Ste20 → Kss1	Fail to correct
GRN8	Flaw: Cdc42 → Ste20	Cdc42 → Ste11	Fail to correct
	Flaw: Ste20 → Ste11	Cdc42 → Ste11	Fail to correct
	Flaw: Ste11→ Ste7	Cdc42 → Ste7	Fail to correct
GRN9	Flaw: Ste7 → Fus3	Ste7 → Far1	Fail to correct
	Flaw: Msg5 → Fus3	Msg5 → Fus3	Succeed in correction
	Flaw: Fus3 → Far1	Fus3 → Far1	Succeed in correction
GRN10	Flaw: Ste7 → Fus3	Ste7 → Fus3	Succeed in correction
	Flaw: Msg5 → Fus3	Msg5 → Fus3	Succeed in correction
	Flaw: Fus3 → Ste12	Msg5→ Ste12	Fail to correct
Accuracy			43%

Figures 4-14 ~ 4-20 show the original networks, defective networks and results of modification, where Modified networks Mod 1~3 are modified from defective networks Def 1~3 respectively by PA.

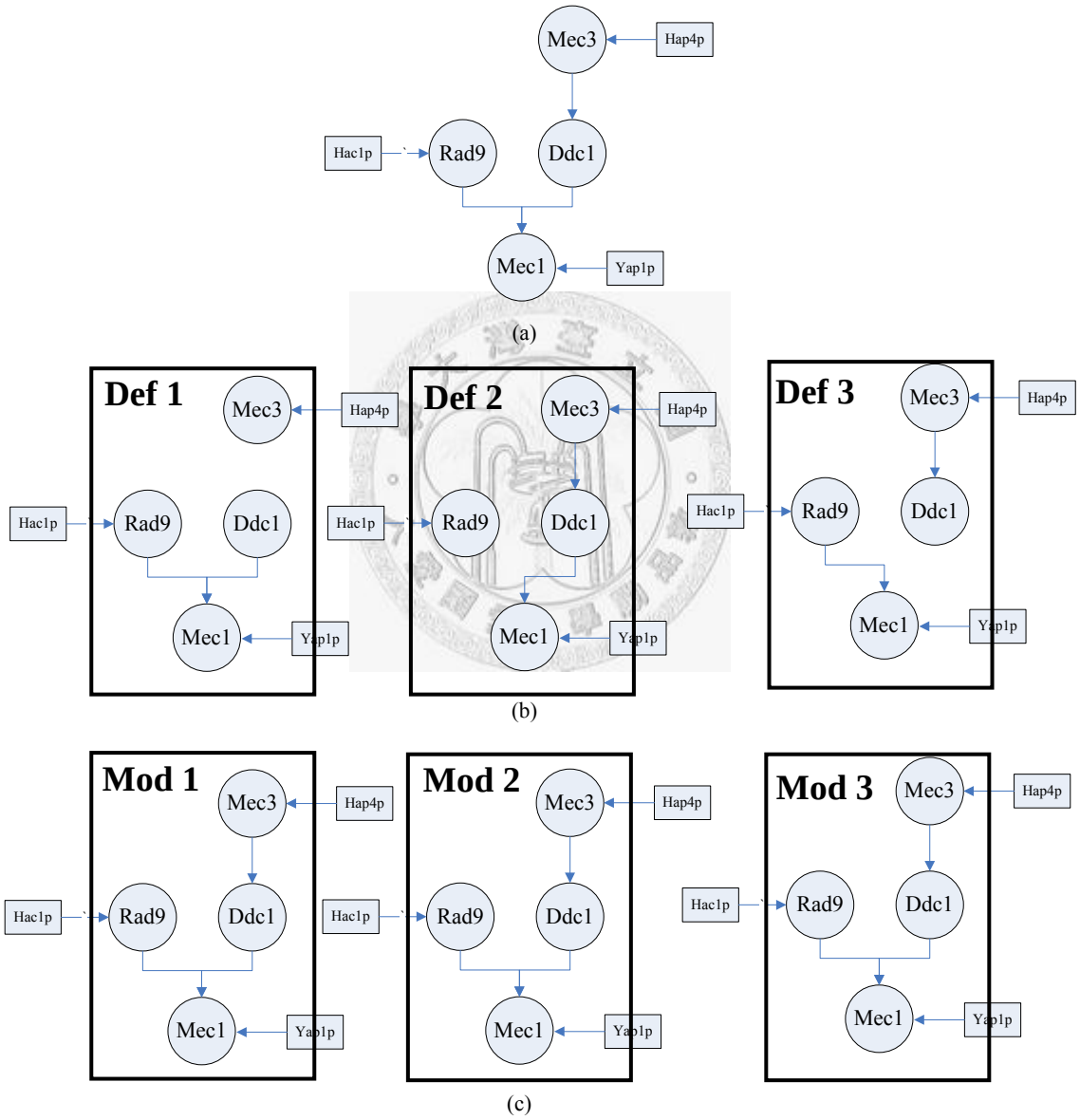


Figure 4-14 GRN2: (a) Original network; (b) Defective networks; (c) Results of modification

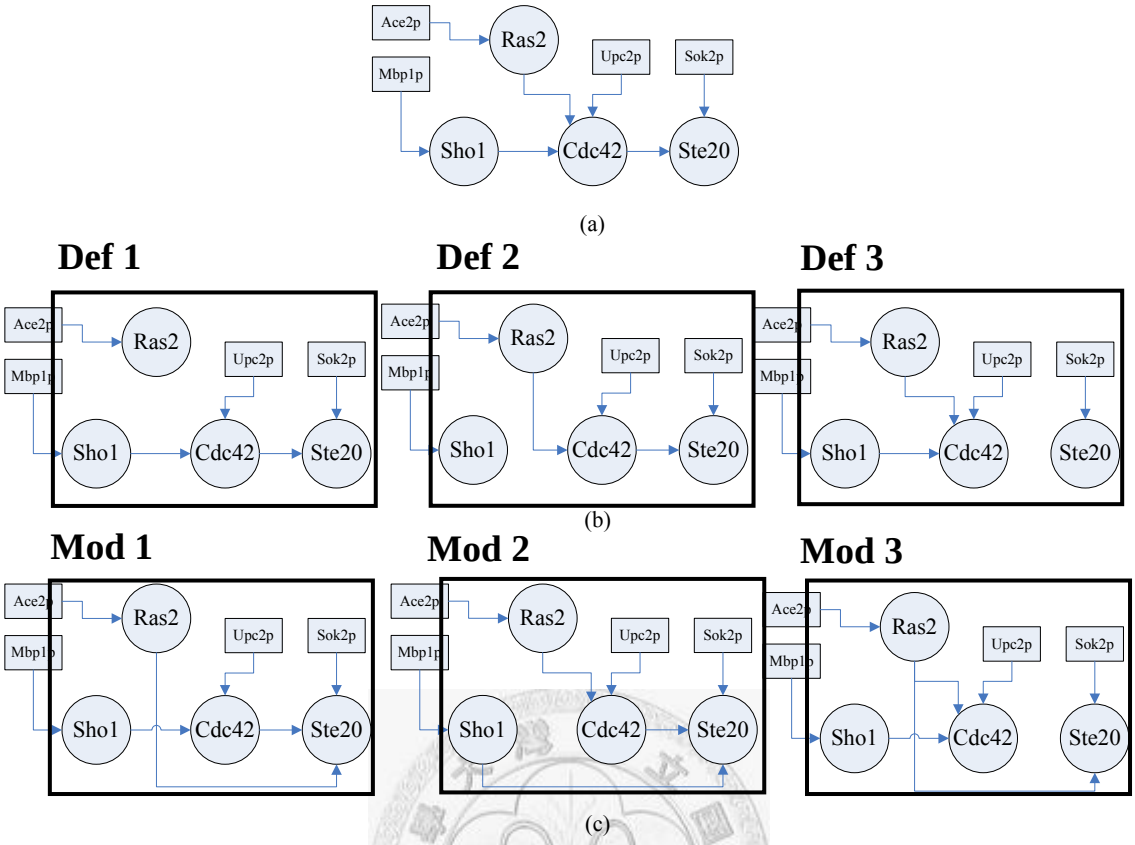


Figure 4-15 GRN5: (a) Original network; (b) Defective networks; (c) Results of modification

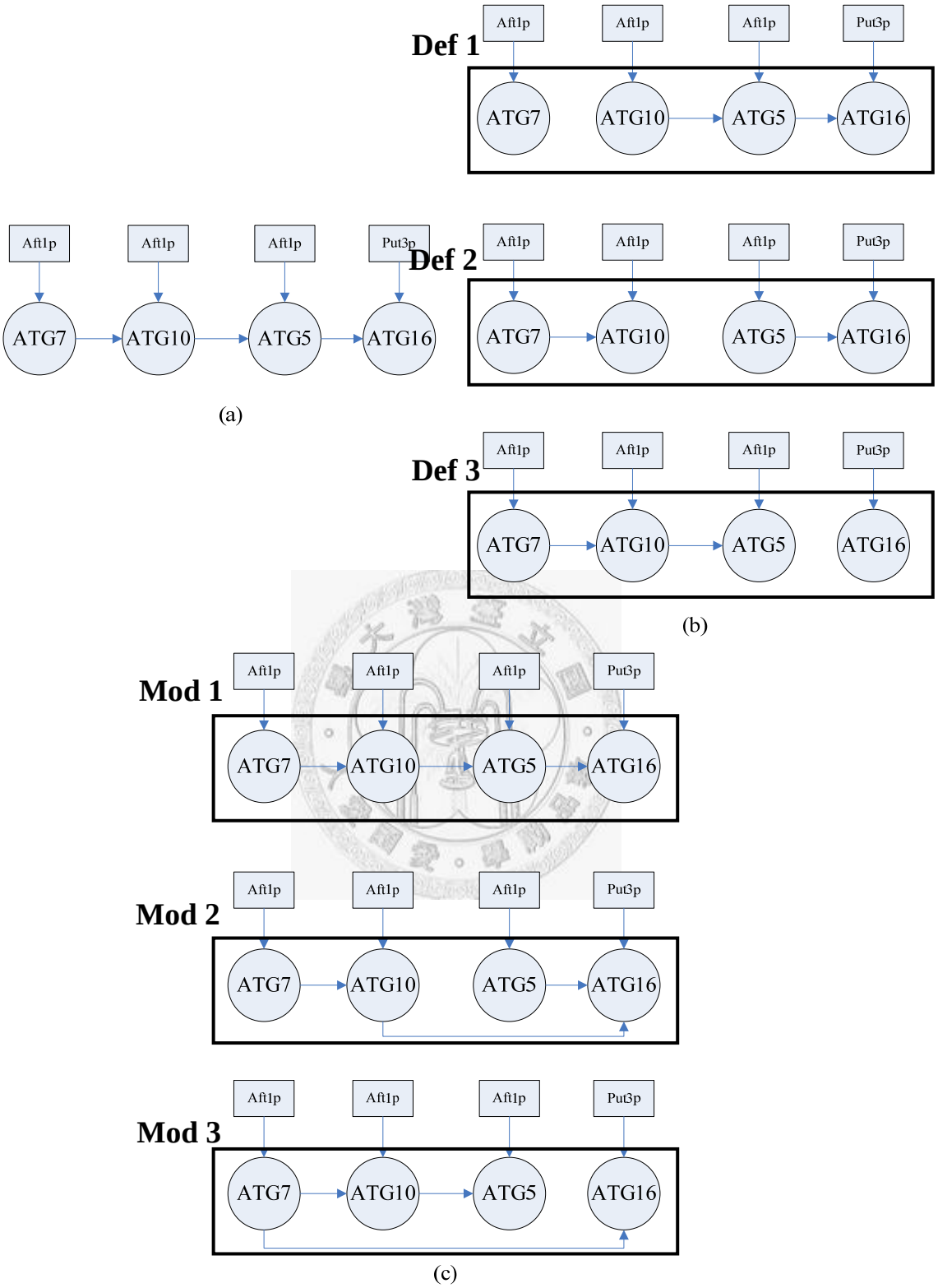


Figure 4-16 GRN6: (a) Original network; (b) Defective networks; (c) Results of modification

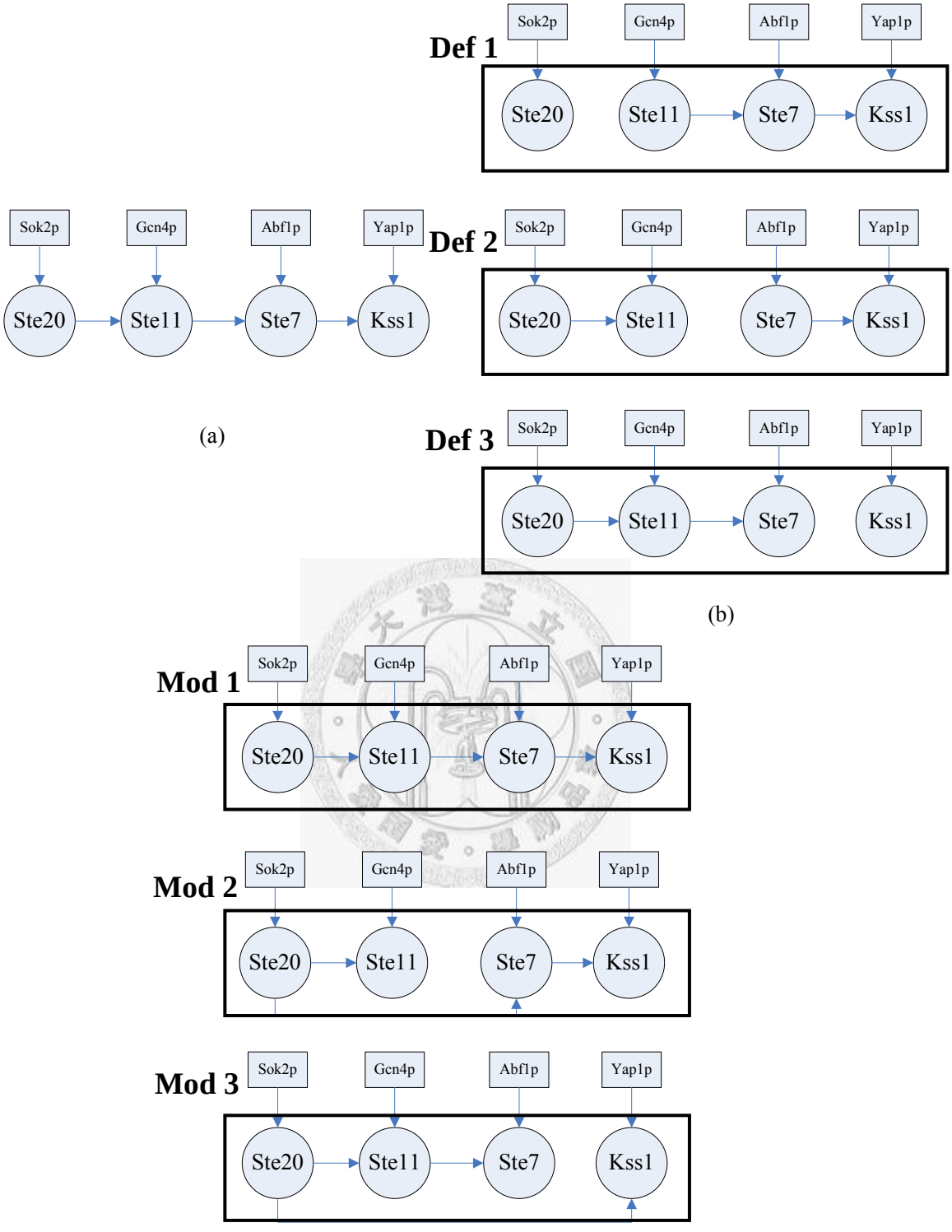


Figure 4-17 GRN7: (a) Original network; (b) Defective networks; (c) Results of modification

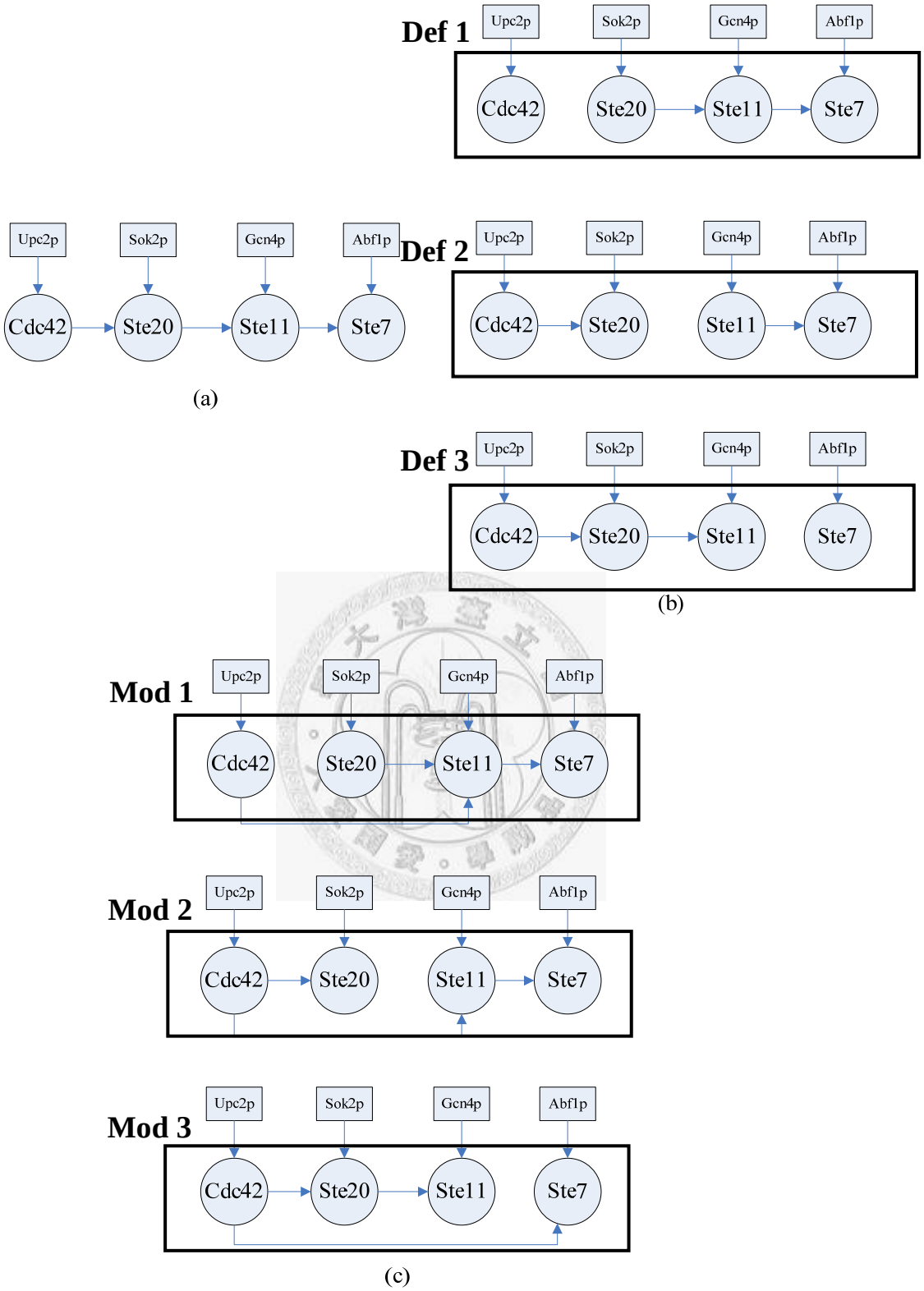


Figure 4-18 GRN8: (a) Original network; (b) Defective networks; (c) Results of modification

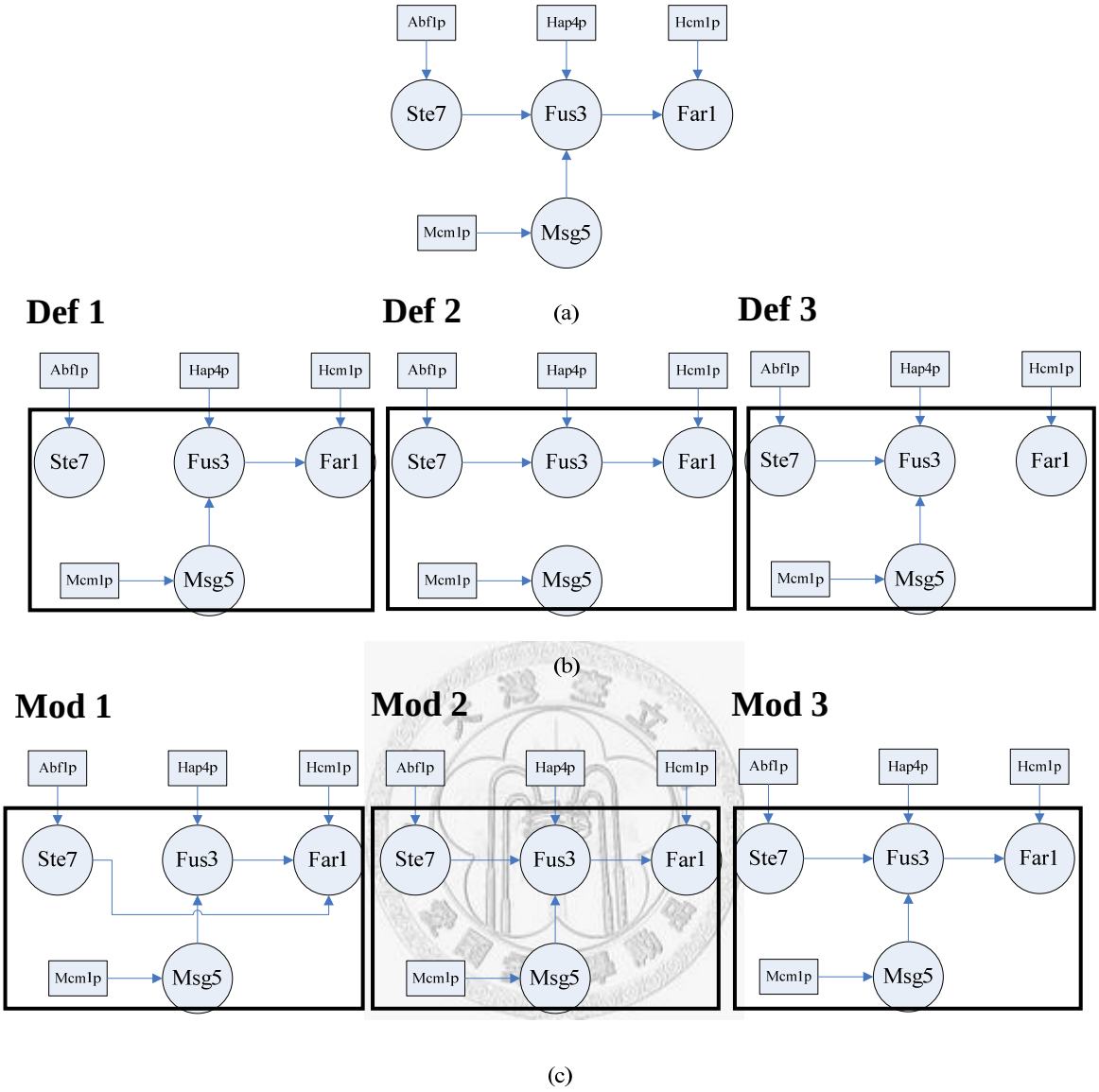


Figure 4-19 GRN9: (a) Original network; (b) Defective networks; (c) Results of modification

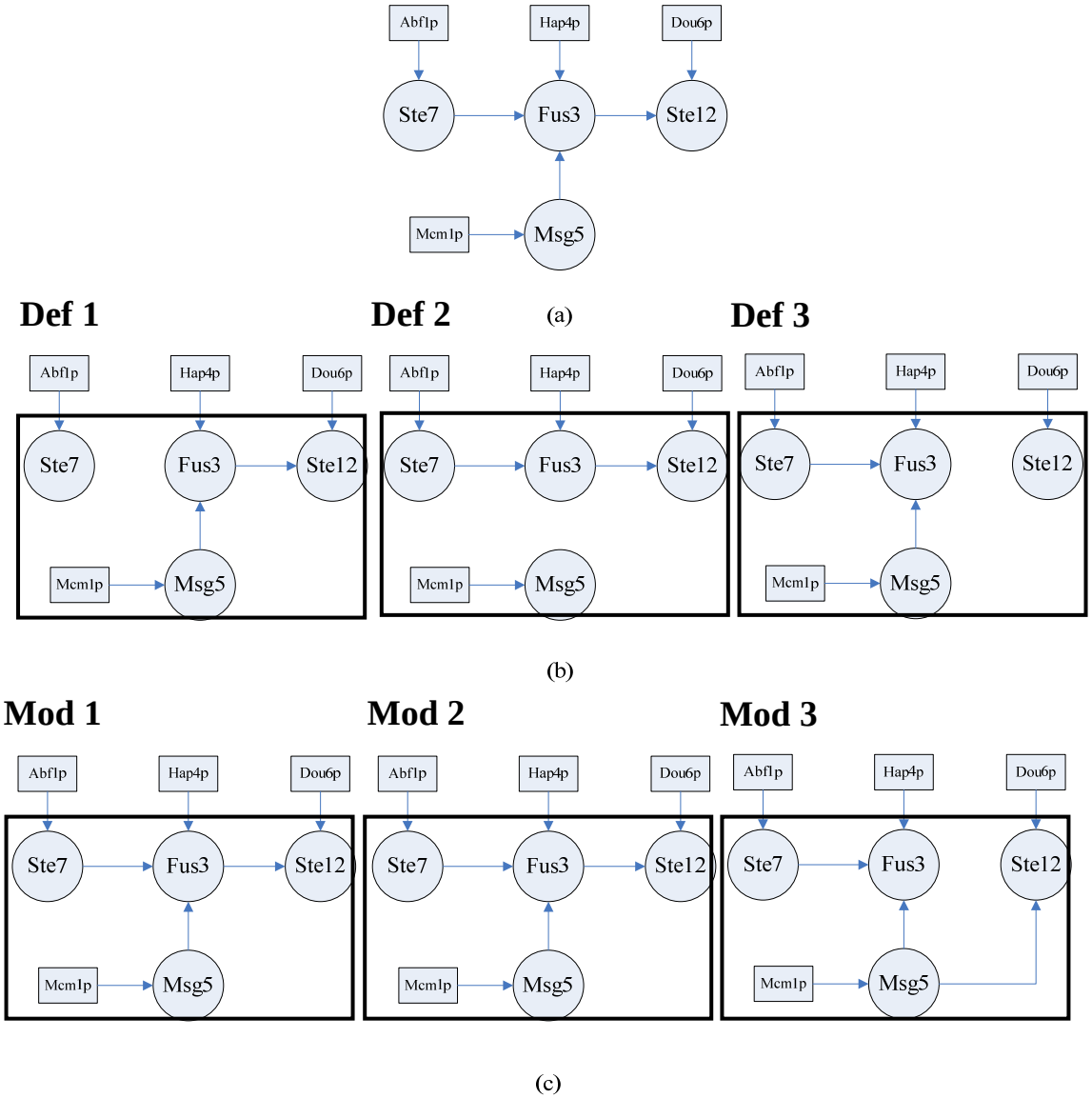


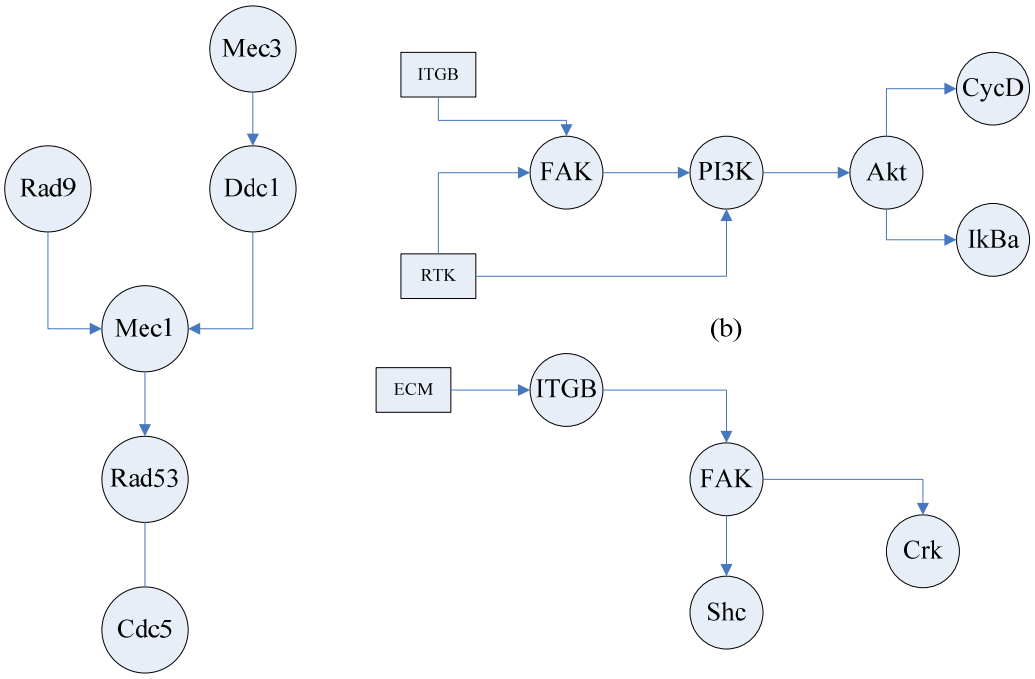
Figure 4-20 GRN10: (a) Original network; (b) Defective networks; (c) Results of modification

4-3 Network extension

In this section, the network which has been identified will be extended to incorporate more genes into the network. We also compare the proposed approach with other algorithms such as SSEM algorithm (Chang *et al*., 2007) and dynamic Bayesian model method (Luna *et al*., 2007).

4-3-1 Comparison with SSEM algorithm

To compare with SSEM algorithm, we use the very three networks that are chosen in their study, as shown in Figure 4-21, for investigation, and these networks are sub-networks regarding yeast cell cycle and Focal adhesion - Homo sapiens (human). Data with 5 time points available in NCBI are adopted by Segal *et al*. (2003) and Duggan *et al*. (2005).



Cell cycle- Saccharomyces cerevisiae **Focal adhesion - Homo sapiens**

Figure 4-21 GRNs in KEGG

4-3-1-1 Comparison with SSEM algorithm on sub-network of cell cycle-yeast

Figure 4-22 shows the sub-network of the network in Figure 4-21 (a), and Table 4-10 shows the results of identification, where original networks wins the majority votes of performance indices in the network pool. Then we add one more gene to the network and generate all possible links for the gene, Rad53. Figure 4-23 shows all possible links (A, B, C, D, E and F) for the new gene, Rad53. Figure 4-24 shows extended networks in Figure 4-23, and Table 4-11 shows the results of identification in Figure 4-24.

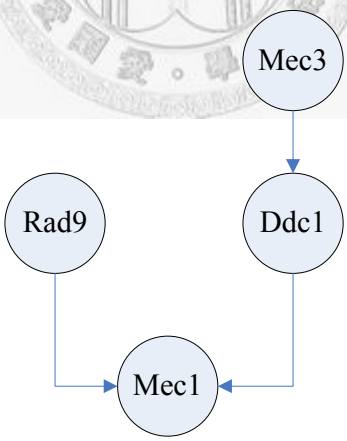


Figure 4-22 Sub-network chosen to be identified in Figure 4-21 (a)

Table 4-10 Results of identification in Figure 4-22

		WLS χ^2	AIC	S-RMR	GFI
Original network		43.23	363.23	0.015	0.77
	1	55.58	375.58	0.230	0.49
	2**				
	3**				
	4	127.6	447.65	0.110	0.33
Alternative networks	5	266.5	586.48	0.180	0.29
	6	44.60	364.60	0.068	0.76
	7	44.77	364.77	0.069	0.76
	8	116.2	436.15	0.410	0.55
	9	44.60	364.60	0.070	0.76

"**": fail in estimation

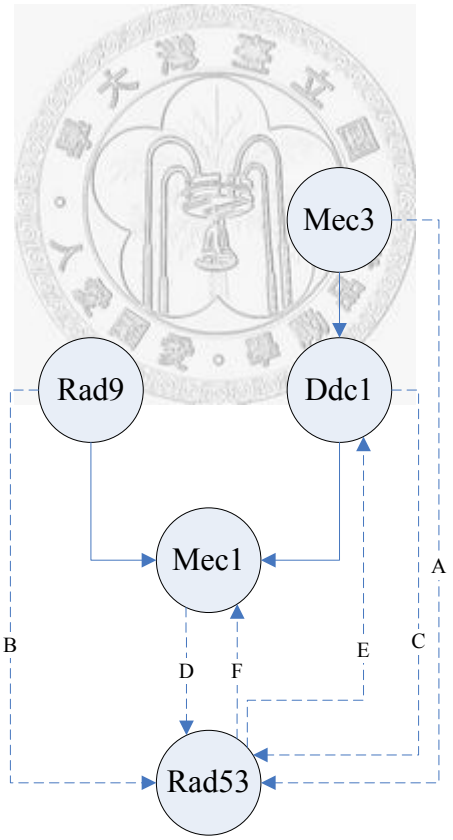


Figure 4-23 All possible links for the new gene, Rad53

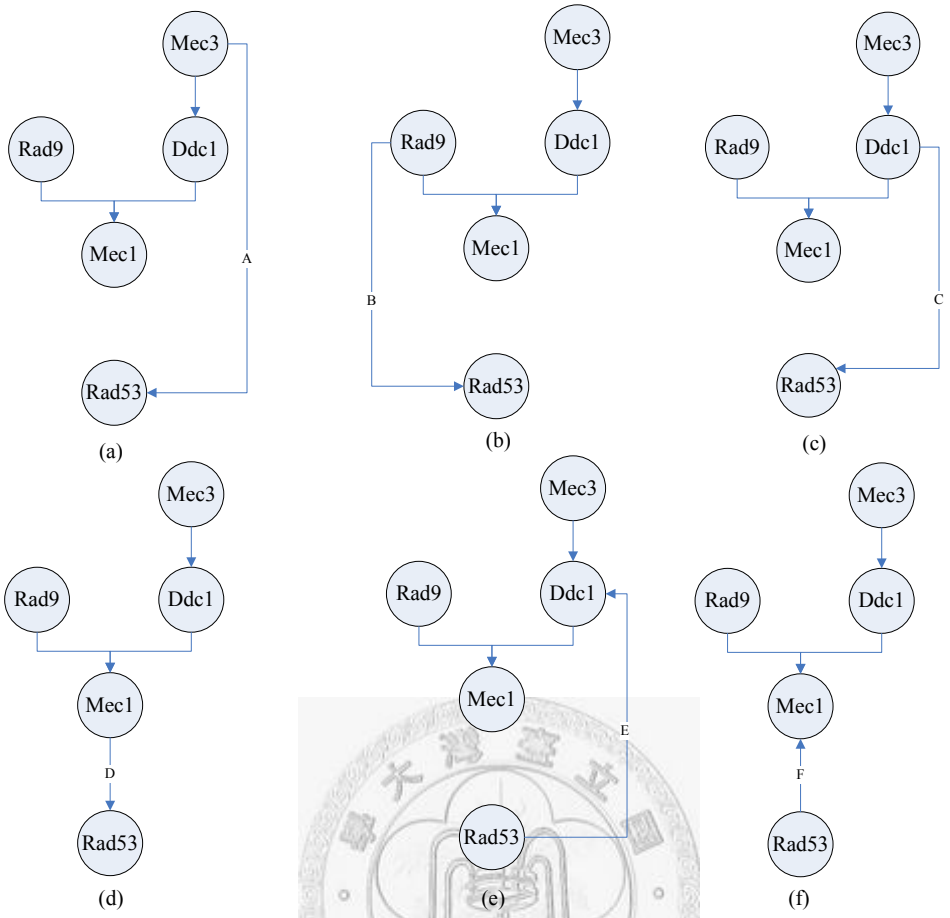


Figure 4-24 Extended networks in Figure 4-23

Table 4-11 Results of identification in Figure 4-24

Extended networks	WLS χ^2	AIC	S-RMR	GFI	Number of performance indices with best scores
(a)**					0/4
(b)	67.39	307.39	0.080	0.73	0/4
(c)	66.25	306.25	0.080	0.73	0/4
(d)	97.61	337.61	0.098	0.59	0/4
(e)	56.68	296.68	0.079	0.76	0/4
(f)	47.62	287.62	0.017	0.79	4/4

“**”: fail in estimation

In Table 4-11, the extended network ‘‘(f)’’ wins the majority votes of performance indices in the network pool, and, we further add one more gene to the network (f) that has been identified and generate all possible links (A, B, C, D, E, F, G and H) for the gene, Cdc5. Figure 4-25 shows the network which has been identified in Figure 4-24 and all possible links for the new gene, Cdc5. Figure 4-26 shows extended networks in Figure 4-25 (b), and Table 4-12 shows the results of identification in Figure 4-26.

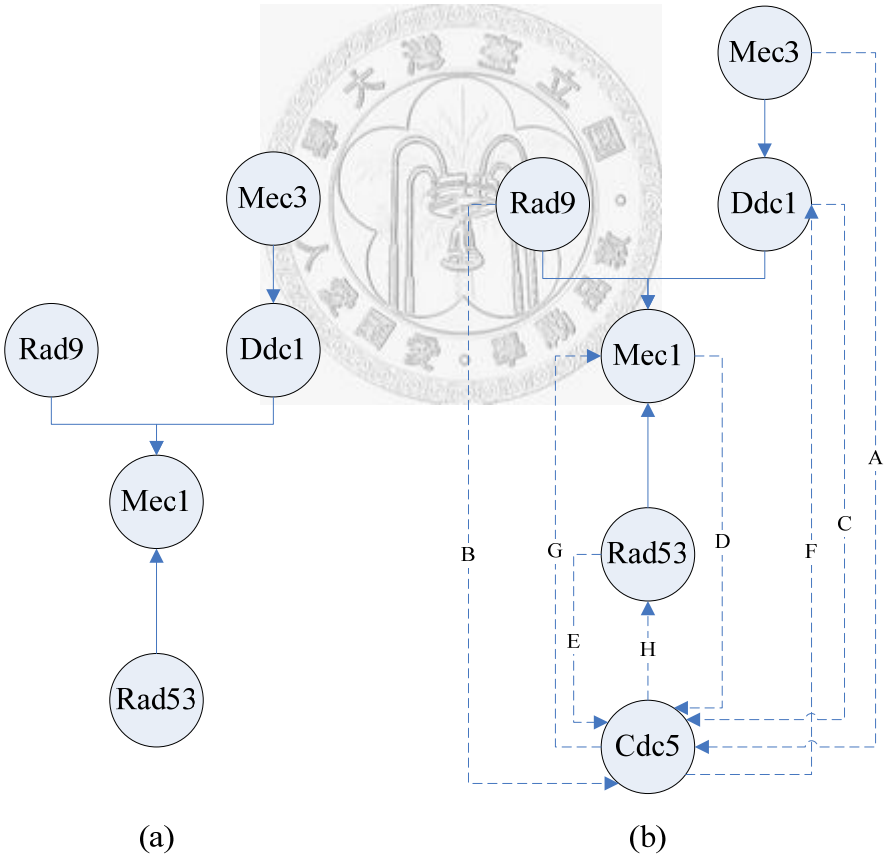


Figure 4-25 (a) Network which has been identified in Figure 4-24; (b) All possible links for the new gene, Cdc5

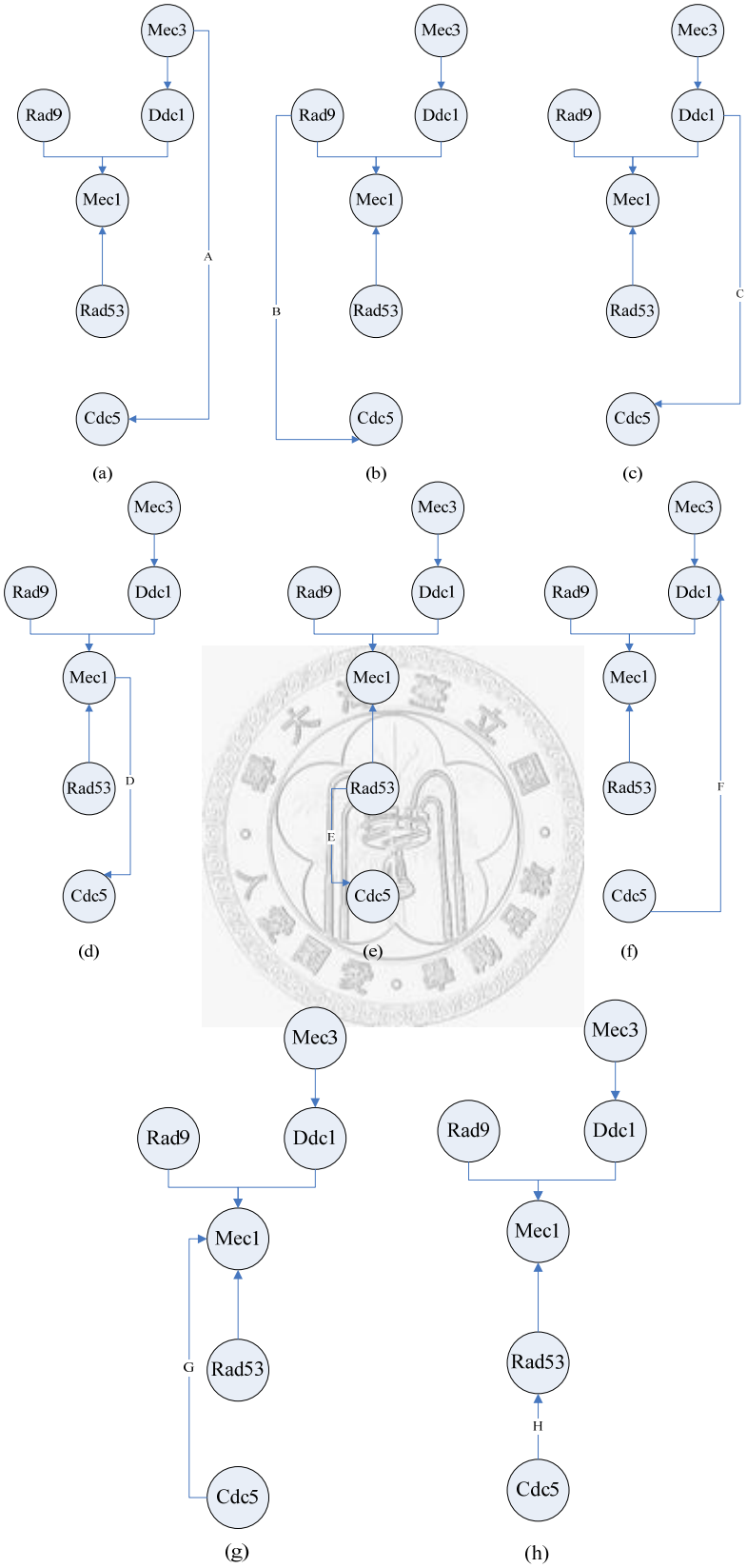


Figure 4-26 Extended networks in Figure 4-25 (b)

Table 4-12 Results of identification in Figure 4-26

Extended networks	WLS χ^2	AIC	S-RMR	GFI	Number of performance indices with best scores
(a)**					0/4
(b)	51.68	331.68	0.013	0.81	1/4
(c)	57.76	337.76	0.013	0.80	1/4
(d)	55.44	335.44	0.013	0.80	1/4
(e)	48.99	328.99	0.013	0.82	4/4
(f)	58.65	338.65	0.013	0.79	1/4
(g)	57.17	337.17	0.013	0.79	1/4
(h)	57.73	337.73	0.013	0.79	1/4

"**": fail in estimation

In Table 4-12, the extended network ''(e)'' wins the majority votes of performance indices in the network pool. Table 4-13 shows the genes and its transcription factors, and Figure 4-27 shows the result of our approach and SSEM algorithm, where the true positive rates are 80% for our approach, and 60% for SSEM algorithm.

Table 4-13 Genes and their transcription factors for Figure 4-21 (a)

Genes	Transcription factors
Mec3	Hap4p
Rad9	Hac1p
Ddc1	
Mec1	Yap1p
Rad53	Arr1p
Cdc5	Mcm1p

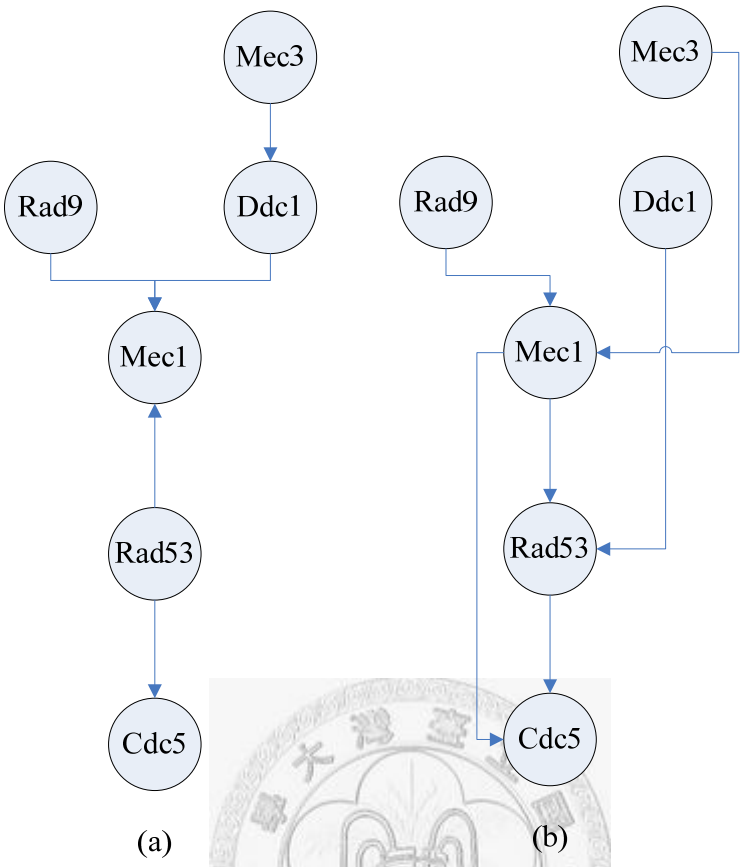


Figure 4-27 (a) Result of our approach; (b) Result of SSEM algorithm

4-3-1-2 Comparison with SSEM algorithm on sub-network of Focal adhesion network (1)

Figure 4-28 shows the sub-network of the network in Figure 4-21 (b), where biological knowledge was considered: RTK affects FAK and PI3K, and ITGB and RTK are exogenous variables (genes). Table 4-14 shows the results of identification, and original networks wins the majority votes of performance indices in the network pool. Then we add one more gene to the network and generate all possible links for the gene, Akt. Figure 4-29 shows all possible links (A, B, C, D, E and F) for the new gene, Akt. Figure 4-30 shows extended networks in Figure 4-29, and Table 4-15 shows the results of identification in Figure 4-30.

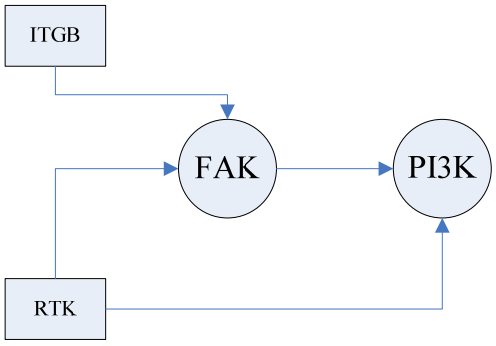


Figure 4-28 Sub-network chosen to identified in Figure 4-21 (b)

Table 4-14 Results of identification in Figure 4-28

		WLS χ^2	AIC	S-RMR	GFI
Original network		16.84	216.84	0.013	0.84
Alternative networks	1	38.23	238.23	0.150	1.00
	2	19.30	219.30	0.007	0.68

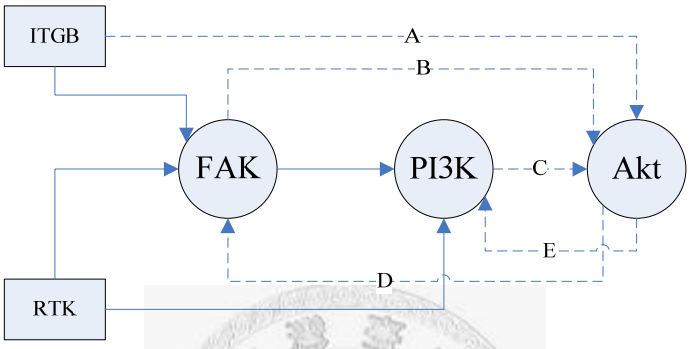


Figure 4-29 All possible links for the new gene, Akt

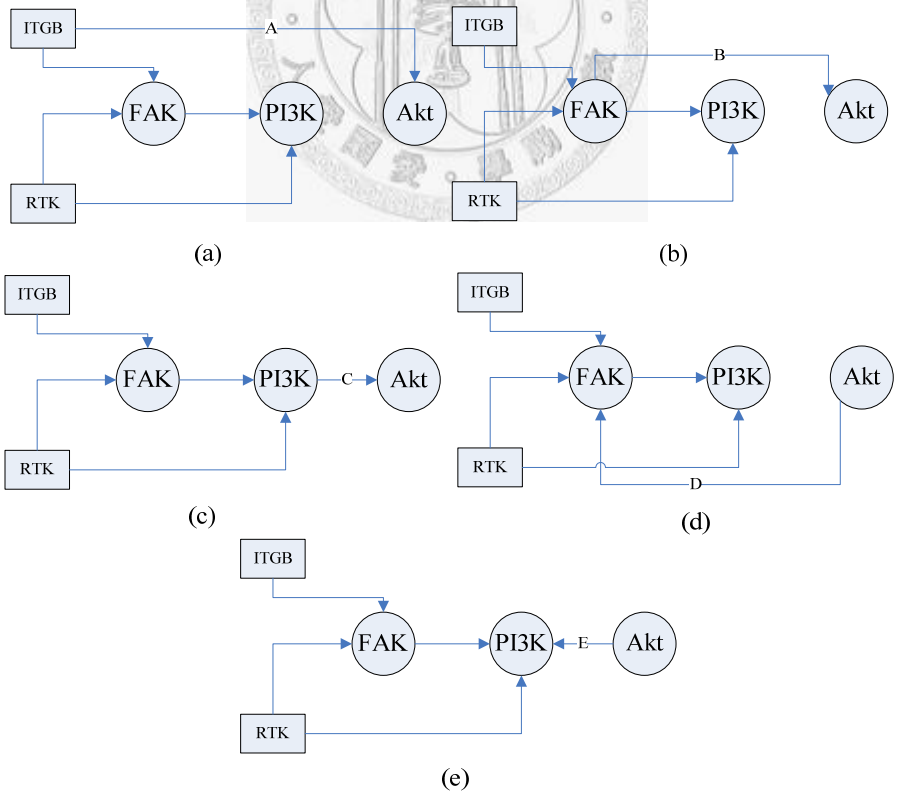


Figure 4-30 Extended networks in Figure 4-29

Table 4-15 Results of identification in Figure 4-30

Extended networks	WLS χ^2	AIC	S-RMR	GFI	Number of performance indices with best scores
(a)**					0/4
(b)	24.24	154.24	0.027	0.81	1/4
(c)	23.38	153.38	0.027	0.82	4/4
(d)	111.06	241.06	0.430	0.48	0/4
(e)	314.54	444.54	0.430	0.32	0/4

"**": fail in estimation

In Table 4-15, the extended network ''(c)'' wins the majority votes of performance indices in the network pool, and further we add one more gene to the network (c) that has been identified and generate all possible links for the gene, CycD. Figure 4-31 shows the network which has been identified in Figure 4-30 and all possible links (A, B, C, D, E and F) for the new gene, CycD. Figure 4-32 shows extended networks in Figure 4-31 (b), and Table 4-16 shows the results of identification in Figure 4-31.

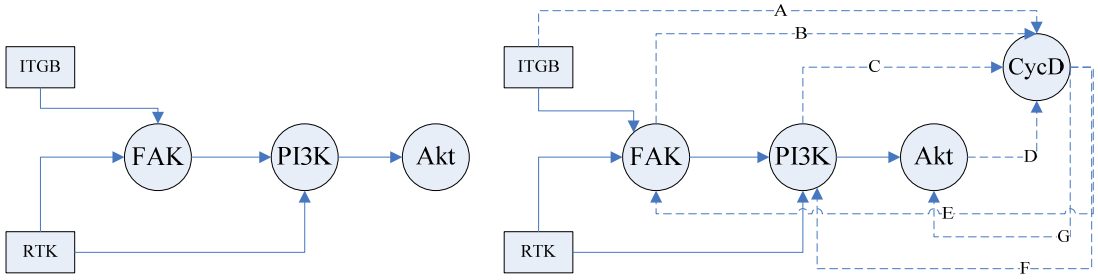


Figure 4-31 (a) Network which has been identified in Figure 4-30; (b) All possible links for the new gene, CycD

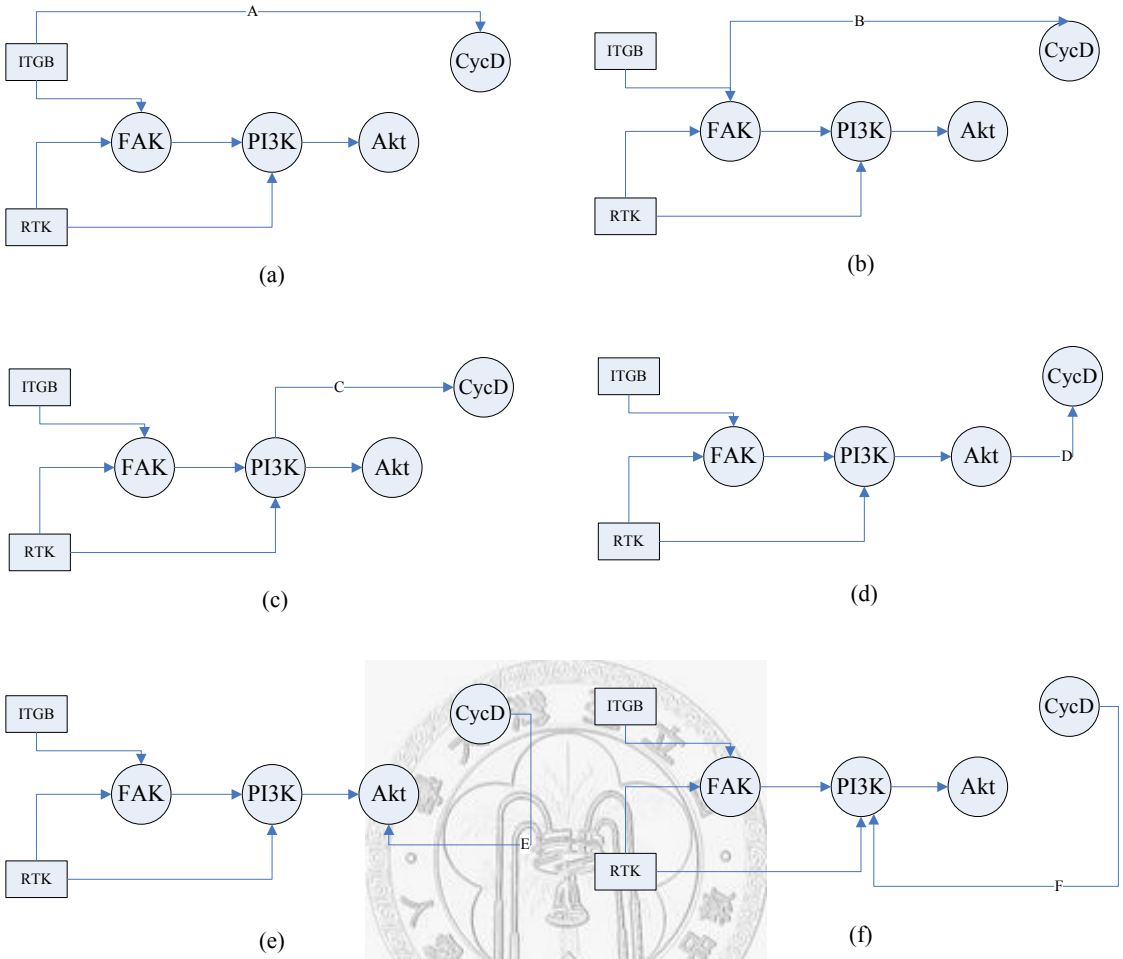


Figure 4-32 Extended networks in Figure 4-31 (b)

Table 4-16 Results of identification in Figure 4-31

Extended networks	WLS χ^2	AIC	S-RMR	GFI	Number of performance indices with best scores
(a)	36.47	186.47	0.210	0.77	0/4
(b)	26.35	176.35	0.023	0.82	4/4
(c)	47.73	197.73	0.046	0.70	0/4
(d)	35.57	185.57	0.023	0.76	1/4
(e)	154.0	304.00	0.430	0.44	0/4
(f)	359.7	509.73	0.440	0.31	0/4

In Table 4-16, the extended network ‘‘(b)’’ wins the majority votes of performance indices in the network pool, and further we add one more gene to the network (b) that has been identified and generate all possible links for the gene, IkBa. Figure 4-33 shows the network which has been identified in Figure 4-32 and all possible links (A, B, C, D, E, F, G and H) for the new gene, IkBa. Figure 4-34 shows extended networks in Figure 4-33 (b), and Table 4-17 shows the results of identification in Figure 4-34.

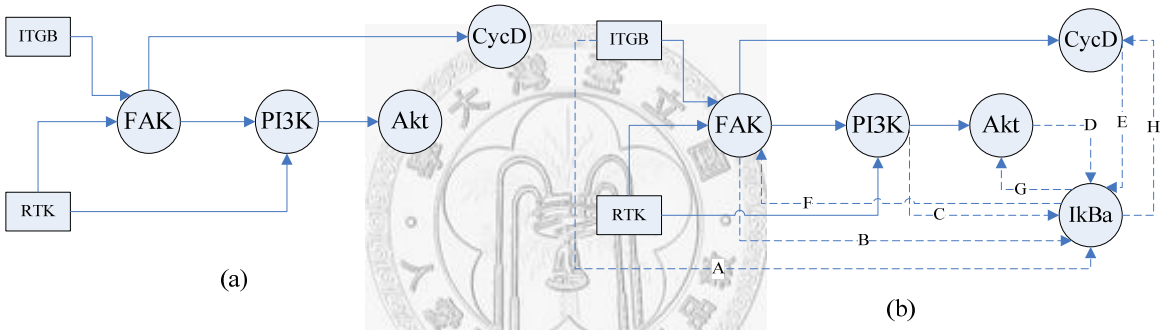


Figure 4-33 (a) Network which has been identified in Figure 4-32; (b) All possible links for the new gene, IkBa

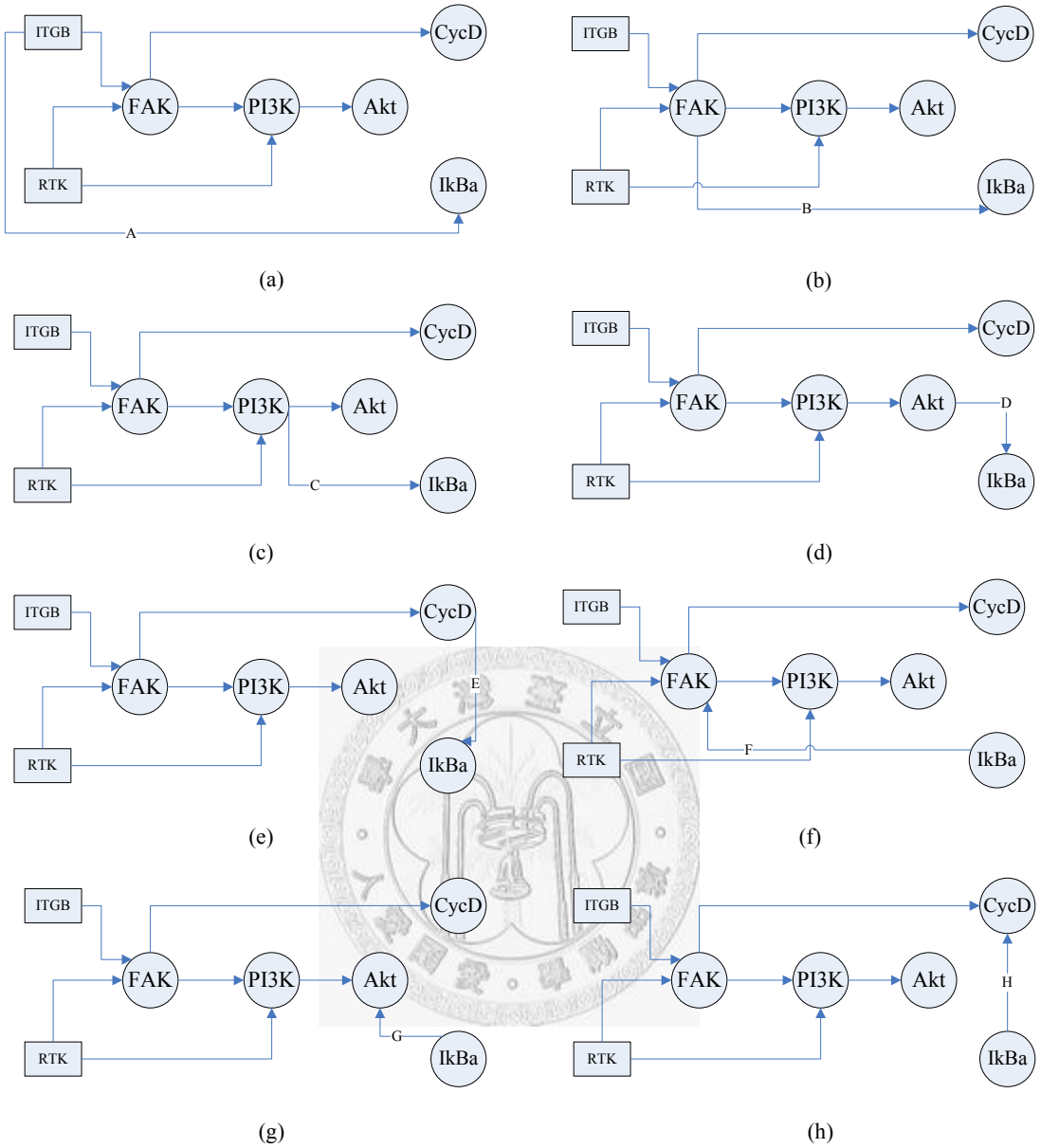


Figure 4-34 Extended networks in Figure 4-33 (b)

Table 4-17 Results of identification in Figure 4-34

Extended networks	WLS χ^2	AIC	S-RMR	GFI	Number of performance indices with best scores
(a)	51.93	221.93	0.023	0.73	0/4
(b)	43.18	213.18	0.022	0.77	1/4
(c)	36.86	206.86	0.022	0.80	4/4
(d)	43.18	213.18	0.022	0.77	1/4
(e)	60.67	230.67	0.022	0.71	0/4
(f)	143.5	313.51	0.400	0.49	0/4
(g)	284.4	454.40	0.400	0.36	0/4
(h)	317.3	387.34	0.400	0.32	0/4

In Table 4-17, the extended network ‘‘(c)’’ wins the majority votes of performance indices in the network pool. Figure 4-35 shows the result of our approach and SSEM algorithm, where the true positive rates are 60% for our approach, and 60% for SSEM algorithm.

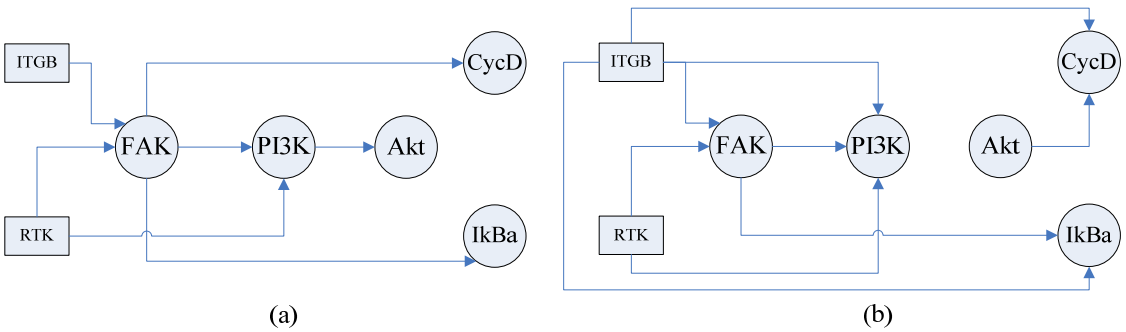


Figure 4-35 (a) Result of our approach; (b) Result of SSEM algorithm

4-3-1-3 Comparison with SSEM algorithm on sub-network of Focal adhesion network (2)

Figure 4-36 shows the sub-network of the network in Figure 4-21 (c), where ECM is exogenous variable (gene). Table 4-18 shows the results of identification, and original networks wins the majority votes of performance indices in the network pool. Then we add one more gene to the network and generate all possible links for the gene, Crk. Figure 4-37 shows all possible links (A, B, C, D and E) for the new gene, Crk. Figure 4-38 shows extended networks in Figure 4-27, and Table 4-19 shows the results of identification in Figure 4-38.

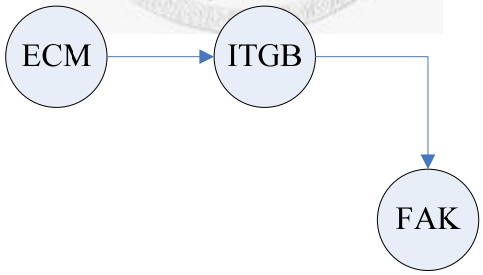


Figure 4-36 Sub-network chosen to identified in Figure 4-21 (c)

Table 4-18 Results of identification in Figure 4-36

		WLS χ^2	AIC	S-RMR	GFI
Original network		0.17	120.17	0.0013	1.00
Alternative networks	1	0.17	120.17	0.0014	1.00
	2	92.4	212.41	0.7200	0.39

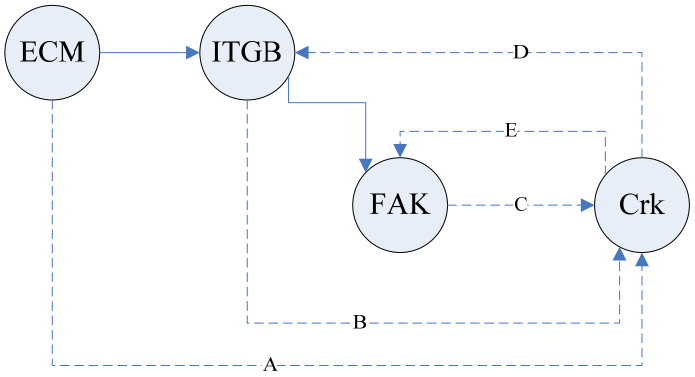


Figure 4-37 All possible links for the new gene, Crk

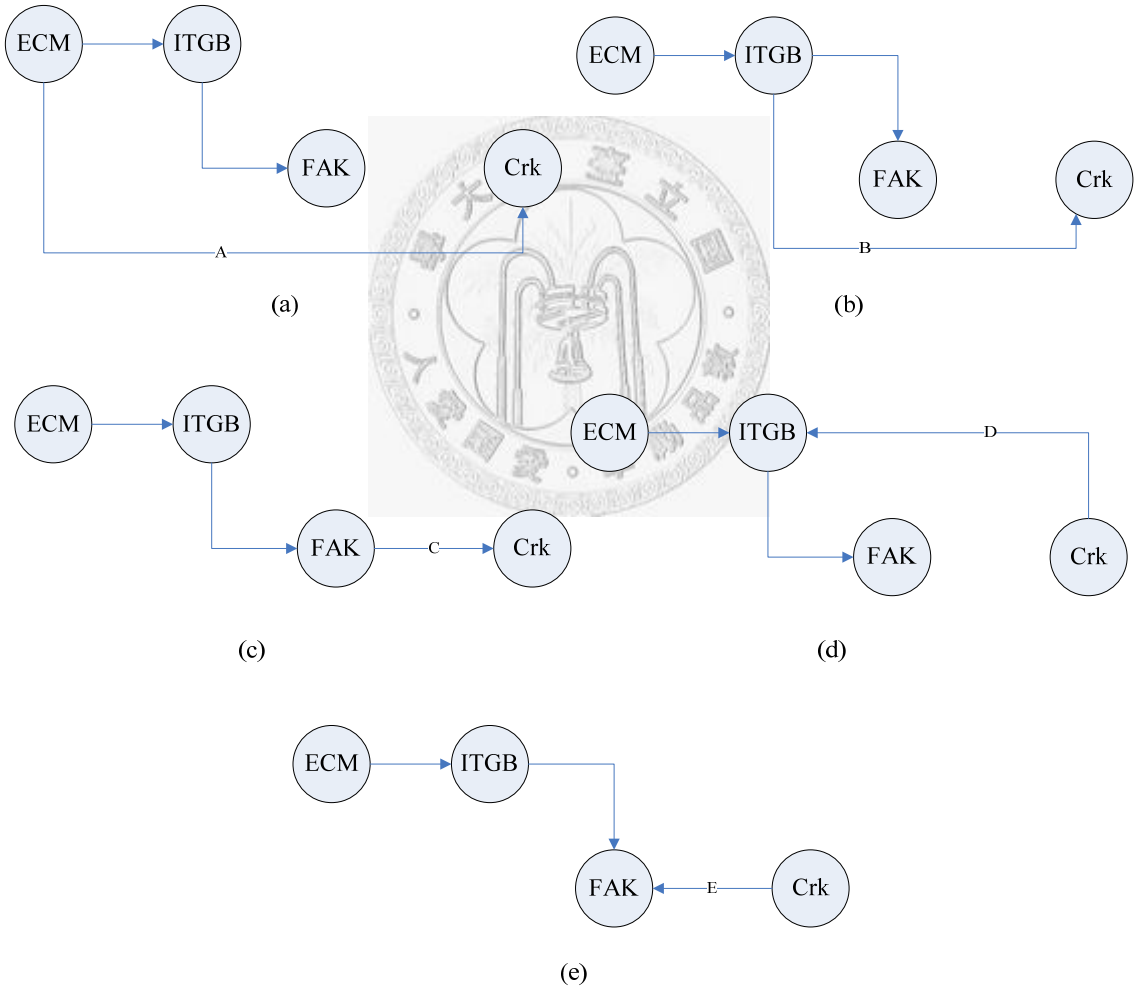


Figure 4-38 Extended networks in Figure 4-37

Table 4-19 Results of identification in Figure 4-38

Extended networks	WLS χ^2	AIC	S-RMR	GFI	Number of performance indices with best scores
(a)**					0/4
(b)	124.6	234.58	0.10	0.48	0/4
(c)	15.09	190.94	0.08	0.85	3/4
(d)	77.16	187.16	0.47	0.52	1/4
(e)	95.62	205.62	0.48	0.46	0/4

"**": fail in estimation

In Table 4-19, the extended network ''(c)'' wins the majority votes of performance indices in the network pool, and further we add one more gene to the network (c) that has been identified and generate all possible links for the gene, Shc. Figure 4-39 shows the network which has been identified in Figure 4-38 and all possible links (A, B, C, D, E, F and G) for the new gene, Shc. Figure 4-40 shows extended networks in Figure 4-39 (b), and Table 4-20 shows the results of identification in Figure 4-40.

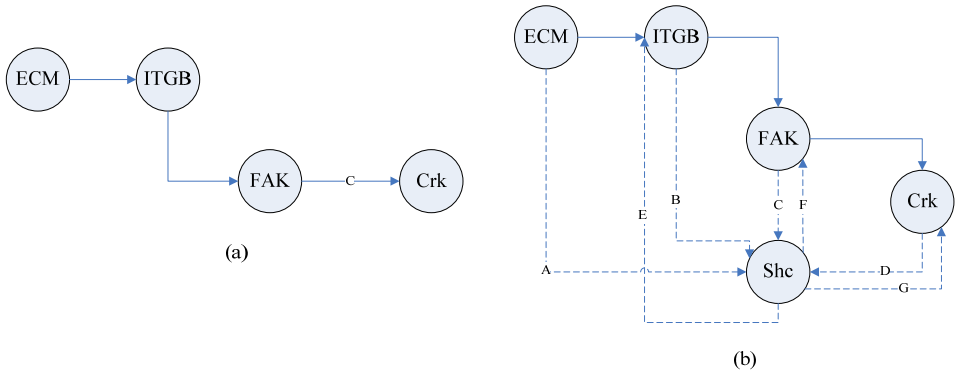


Figure 4-39 (a) Network which has been identified in Figure 4-38; (b) All possible links for the new gene, Shc

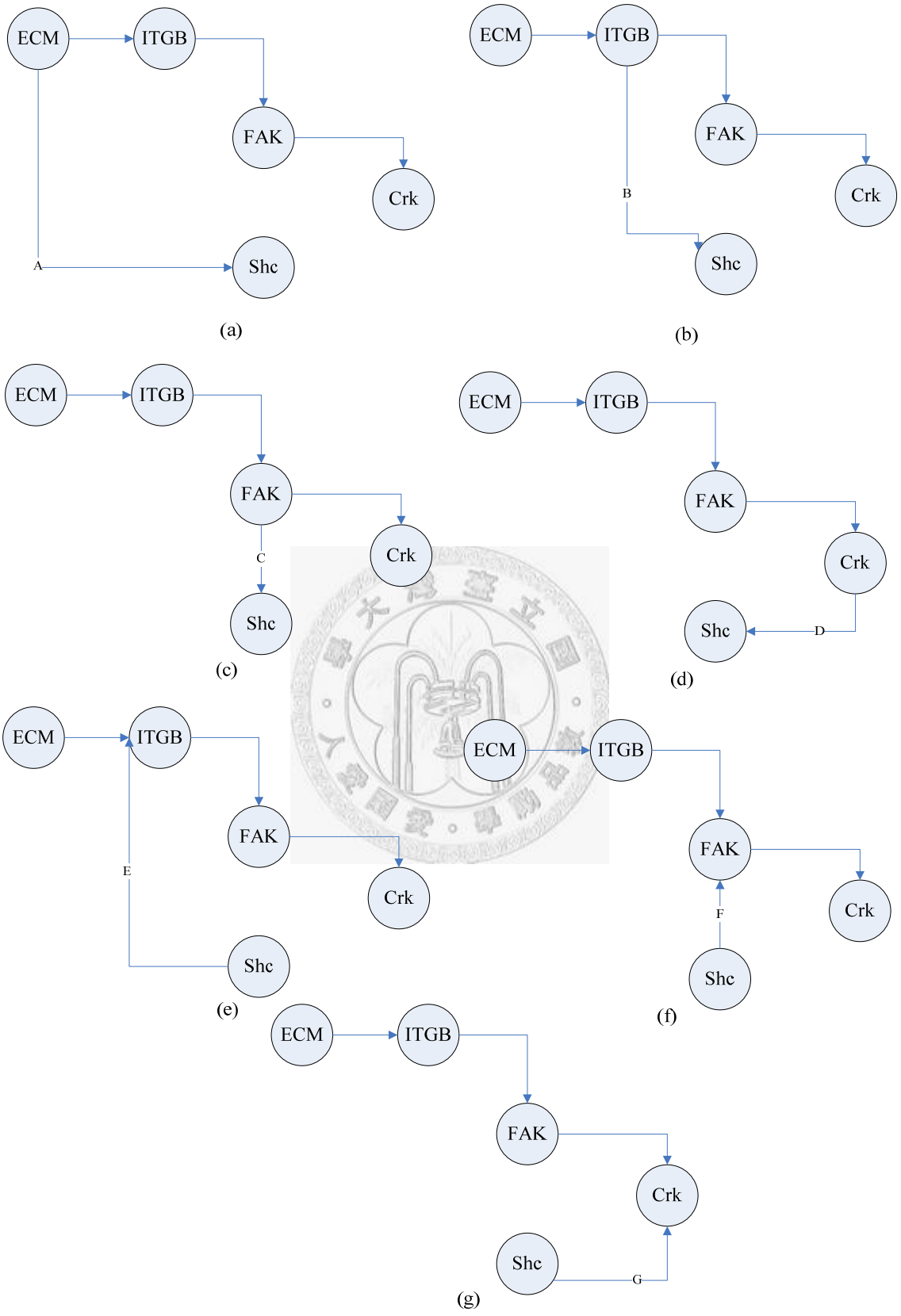


Figure 4-40 Extended networks in Figure 4-39 (b)

Table 4-20 Results of identification in Figure 4-40

Extended networks	WLS χ^2	AIC	S-RMR	GFI	Number of performance indices with best scores
(a)	44.66	174.66	0.170	0.52	0/4
(b)	26.65	156.65	0.077	0.80	0/4
(c)	23.18	153.18	0.075	0.82	4/4
(d)	31.81	161.81	0.076	0.77	0/4
(e)	122.3	252.25	0.490	0.45	0/4
(f)	169.3	299.29	0.490	0.35	0/4
(g)	128.1	158.13	0.500	0.45	0/4

In Table 4-20, the extended network ‘‘(c)’’ wins the majority votes of performance indices in the network pool. Figure 4-41 shows the result of our approach and SSEM algorithm, where the true positive rates are 100% for our approach, and 75 % for SSEM algorithm.

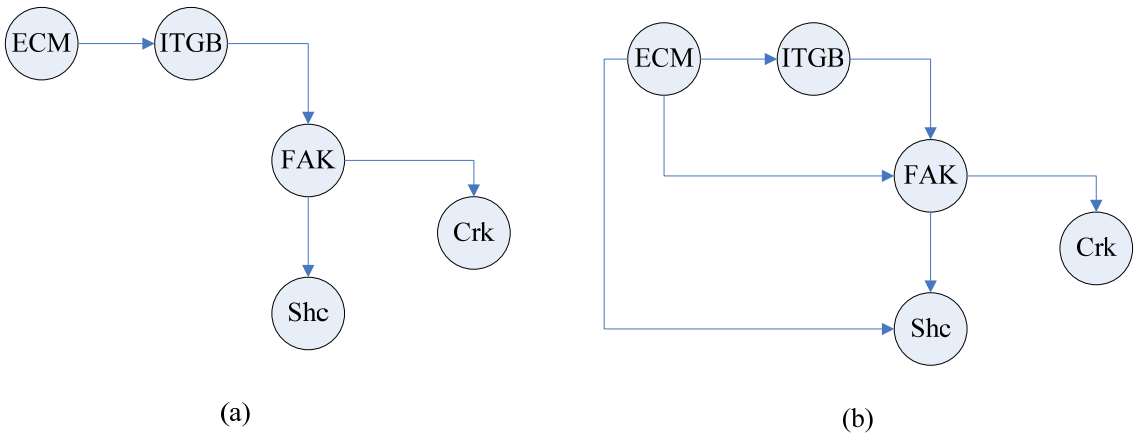


Figure 4-41 (a) Result of our approach; (b) Result of SSEM algorithm

4-3-2 Comparison with dynamic Bayesian model method

To compare with dynamic Bayesian model method (Luna *et al.*, 2007), we use the one network that is chosen in their study, as shown in Figure 4-42, for investigation, and this network is sub-networks regarding yeast cell cycle. Alpha data sets that are also chosen in their study with 6 time points are adopted by Spellman *et al*. (1998).

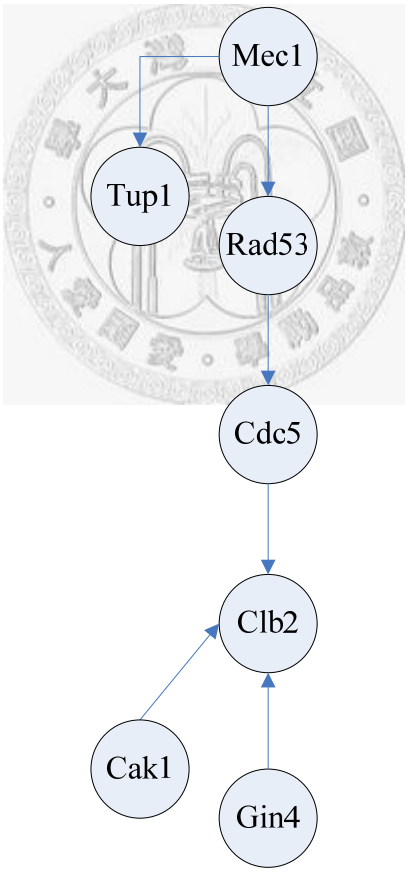


Figure 4-42 GRNs in KEGG

Figure 4-43 shows the sub-network of the network as shown in Figure 4-42, and Table 4-21 shows the results of identification, where original networks wins the majority votes of performance indices in the network pool. Then we add one more gene to the network and generate all possible links for the gene, Cdc5. Figure 4-44 shows all possible links (A, B, C, D, E, F and G) for the new gene, Cdc5. Figure 4-45 shows extended networks in Figure 4-44, and Table 4-22 shows the results of identification in Figure 4-45.

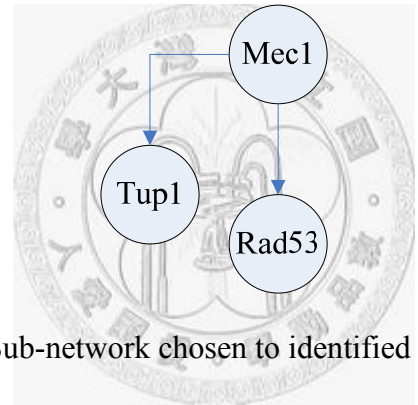


Figure 4-43 Sub-network chosen to identified in Figure 4-42

Table 4-21 Results of identification in Figure 4-43

		WLS χ^2	AIC	S-RMR	GFI
Original network		39.46	393.46	0.012	0.80
Alternative networks	1	45.43	399.43	0.012	0.78
	2	64.69	418.69	0.250	0.43

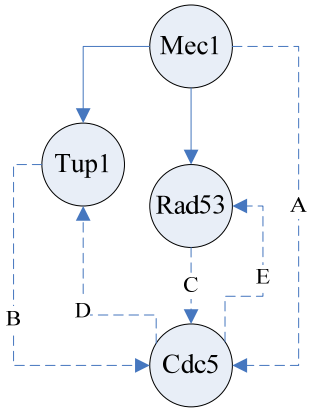


Figure 4-44 All possible links for the new gene, Cdc5

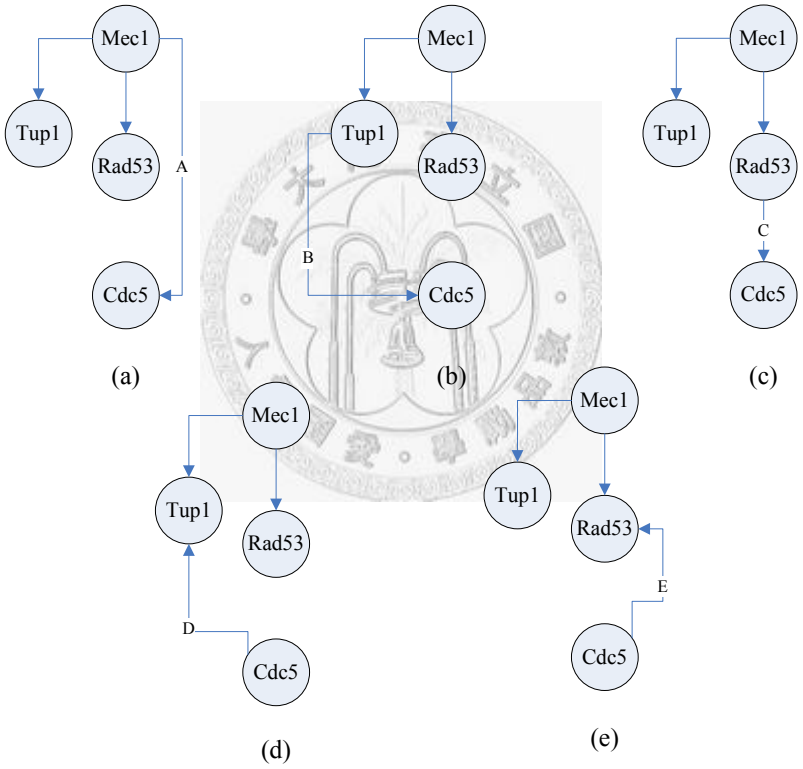


Figure 4-45 Extended networks in Figure 4-44

Table 4-22 Results of identification in Figure 4-45

Extended networks	WLS χ^2	AIC	S-RMR	GFI	Number of performance indices with best scores
(a)**					0/4
(b)	70.14	346.14	0.017	0.75	1/4
(c)	63.01	339.01	0.017	0.76	4/4
(d)	79.61	355.61	0.019	0.72	0/4
(e)	75.54	351.54	0.017	0.73	1/4

"**": fail in estimation

In Table 4-22, the extended network ''(c)'' wins the majority votes of performance indices in the network pool, and further we add one more gene to the network (c) that has been identified and generate all possible links for the gene, Clb2. Figure 4-46 shows the network which has been identified in Figure 4-45 and all possible links (A, B, C, D, E, F and G) for the new gene, Clb2. Figure 4-47 shows extended networks in Figure 4-46 (b), and Table 4-23 shows the results of identification in Figure 4-47.

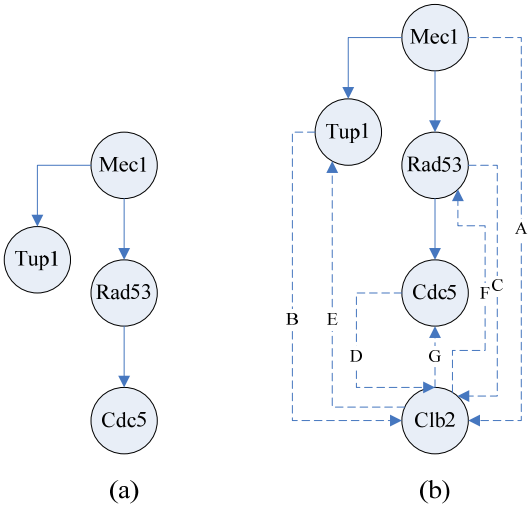


Figure 4-46 (a) Network which has been identified in Figure 4-45; (b) All possible links for the new gene, Clb2

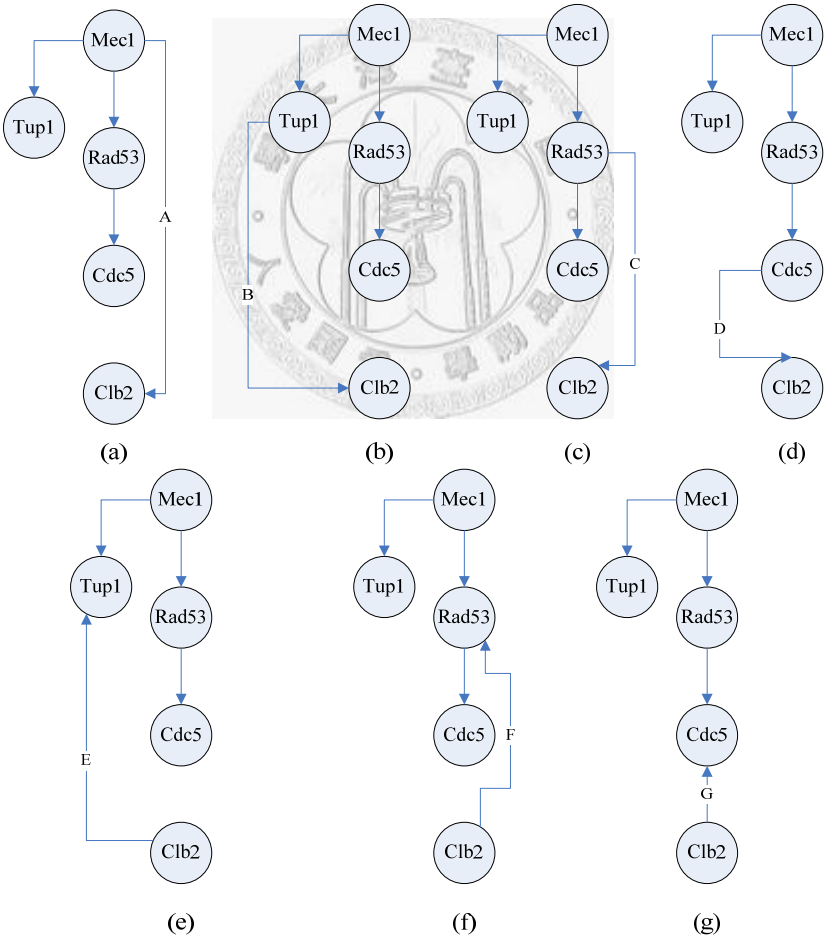


Figure 4-47 Extended networks in Figure 4-46 (b)

Table 4-23 Results of identification in Figure 4-47

Extended networks	WLS χ^2	AIC	S-RMR	GFI	Number of performance indices with best scores
(a)	272.4	596.4	0.26	0.42	1/4
(b)	330.3	657.3	0.27	0.41	0/4
(c)	281.2	605.2	0.26	0.41	1/4
(d)	279.1	603.1	0.31	0.48	0/4
(e)	427.9	751.9	0.29	0.34	0/4
(f)	408.2	732.2	0.29	0.34	0/4
(g)	133.2	457.2	0.29	0.66	4/4

In Table 4-23, the extended network ‘‘(g)’’ wins the majority votes of performance indices in the network pool, and further we add one more gene to the network (g) that has been identified and generate all possible links for the gene, Cak1. Figure 4-48 shows the network which has been identified in Figure 4-47 and all possible links (A, B, C, D, E, F, G, H and I) for the new gene, Cak1. Figure 4-49 shows extended networks in Figure 4-48 (b), and Table 4-24 shows the results of identification in Figure 4-49. Because covariance matrix is not positive-definite, ridge constant is used to make correction, LISREL 8.8.

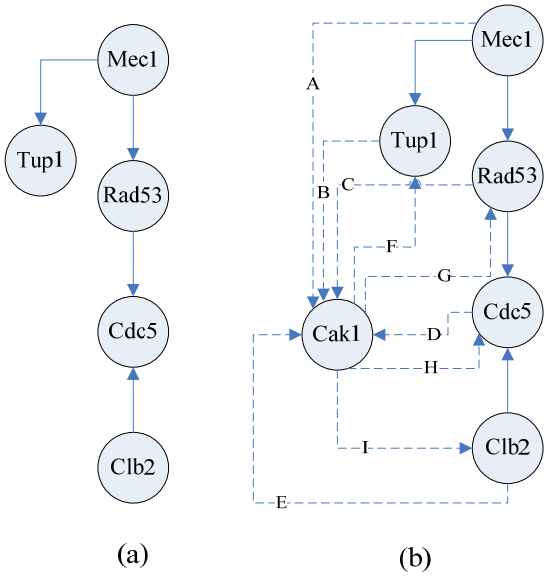


Figure 4-48 (a) Network which has been identified in Figure 4-47; (b) All possible links for the new gene, Cak1

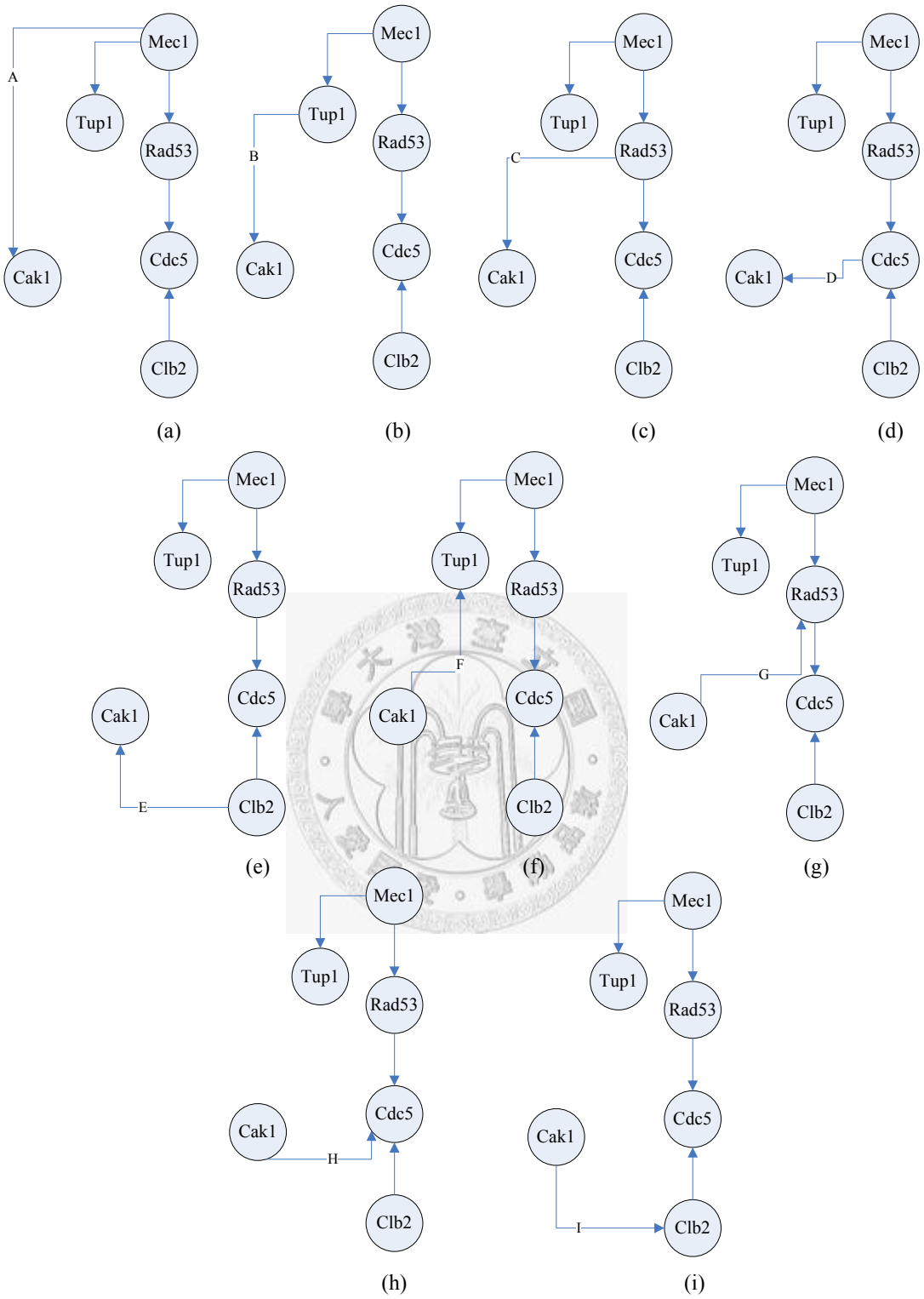


Figure 4-49 Extended networks in Figure 4-48 (b)

Table 4-24 Results of identification in Figure 4-49

Extended networks	WLS χ^2	AIC	S-RMR	GFI	Number of performance indices with best scores
(a)	199.5	571.52	0.25	0.60	0/4
(b)	199.5	571.54	0.25	0.60	0/4
(c)	199.5	571.53	0.25	0.60	0/4
(d)	253.1	625.10	0.23	0.51	0/4
(e)	542.9	914.86	0.23	0.35	0/4
(f)	202.1	574.13	0.25	0.60	0/4
(g)	311.8	683.77	0.26	0.51	0/4
(h)	211.4	583.42	0.25	0.61	0/4
(i)	22.30	394.25	0.07	0.93	4/4

In Table 4-24, the extended network ‘‘(i)’’ wins the majority votes of performance indices in the network pool, and further we add one more gene to the network (i) that has been identified and generate all possible links for the gene, Gin4. Figure 4-50 shows the network which has been identified in Figure 4-49 and all possible links (A, B, C, D, E, F, G, H, I, J and K) for the new gene, Gin4. Figure 4-51 shows extended networks in Figure 4-50 (b), and Table 4-25 shows the results of identification in Figure 4-51. Because covariance matrix is not positive-definite, ridge constant is used to make correction, LISREL 8.8.

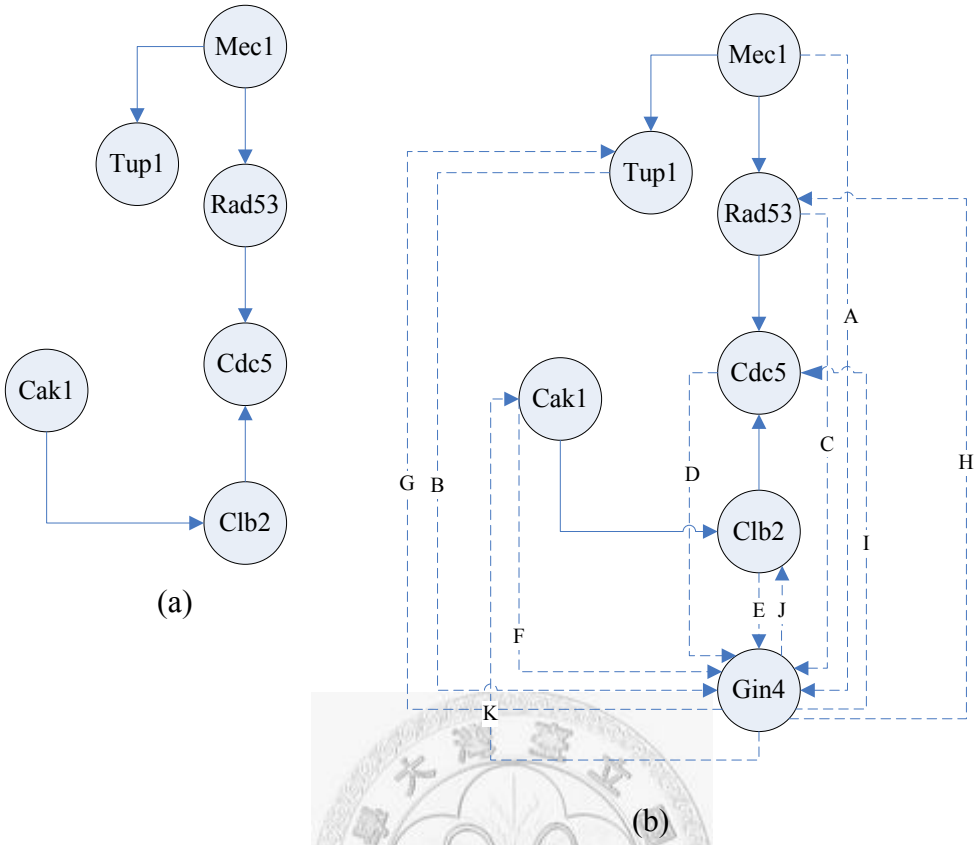


Figure 4-50 (a) Network which has been identified in Figure 4-49; (b) All possible links for the new gene, Gin4

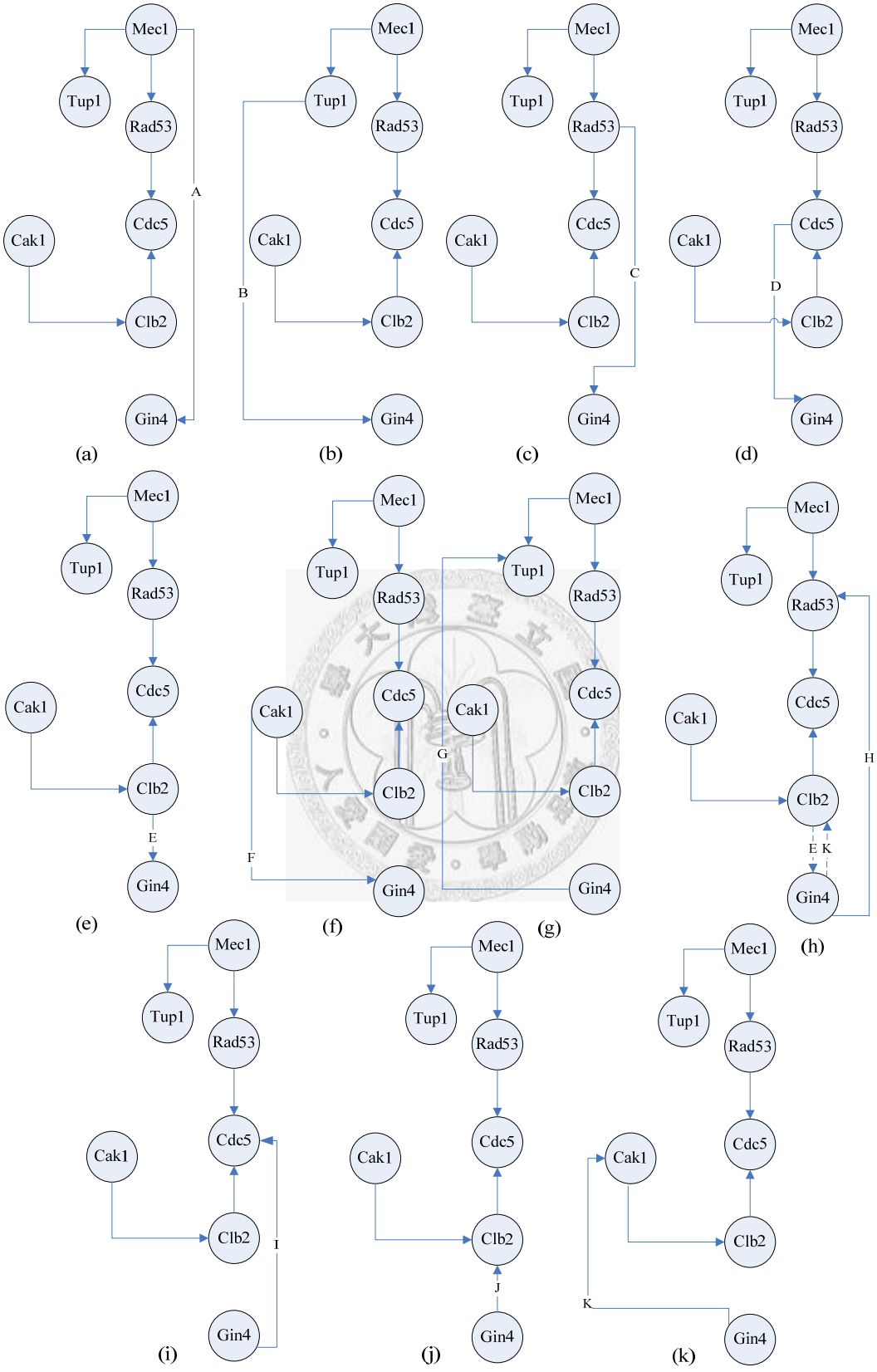


Figure 4-51 Extended networks in Figure 4-50 (b)

Table 4-25 Results of identification in Figure 4-51

Extended networks	WLS χ^2	AIC	S-RMR	GFI	Number of performance indices with best scores
(a)	35.68	455.68	0.11	0.91	0/4
(b)	35.68	455.68	0.11	0.91	0/4
(c)	78.37	498.37	0.13	0.77	0/4
(d)	69.59	489.59	0.13	0.78	0/4
(e)	72.42	492.42	0.12	0.79	0/4
(f)	35.77	455.77	0.11	0.91	0/4
(g)	35.67	455.67	0.11	0.91	0/4
(h)	35.17	455.17	0.10	0.91	0/4
(i)	25.52	445.52	0.09	0.93	0/4
(j)	22.05	442.05	0.07	0.94	4/4
(k)	22.52	442.52	0.07	0.94	2/4

In Table 4-25, the extended network ‘‘(j)’’ wins the majority votes of performance indices in the network pool. Figure 4-52 shows the result of our approach and dynamic Bayesian model method, where the true positive rates are 83% for our approach. Our approach has better true positive rates than dynamic Bayesian model method, but the network that has been identified is small than dynamic Bayesian model method.

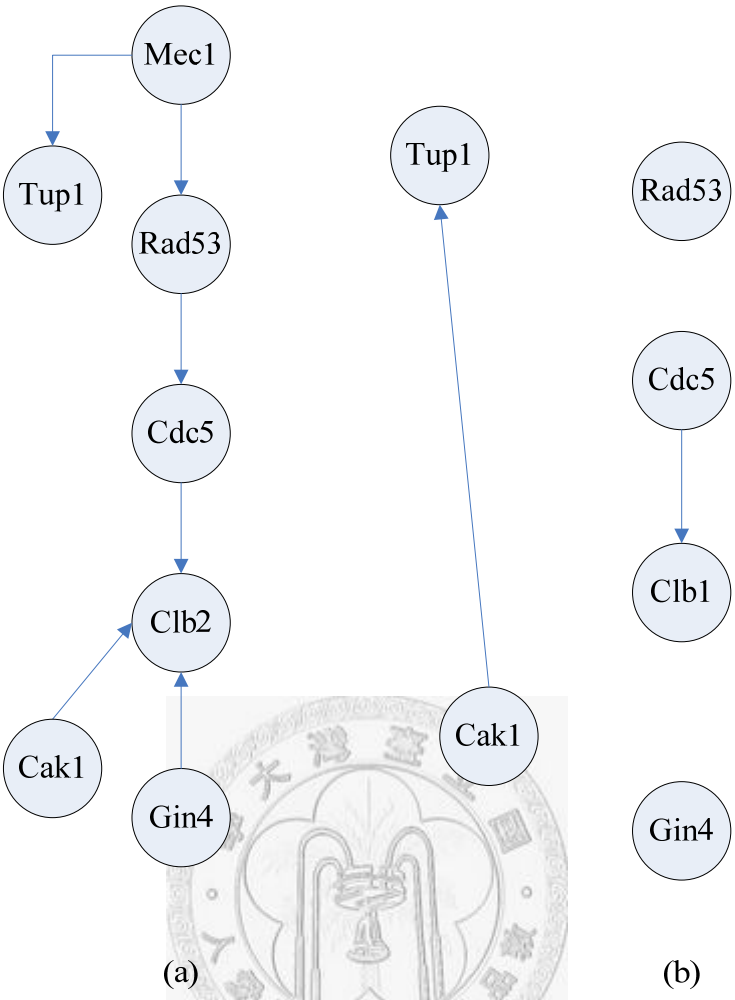


Figure 4-52 (a) Result of our approach; (b) Result of dynamic Bayesian model method

Chapter 5 Conclusions

We proposed the novel identification and modification approach to identify gene regulatory networks and to evaluate the ability of correction for defective networks by dynamic path analysis (PA) algorithm. In ten original gene regulatory networks from KEGG, including sub-networks of cell cycle –yeast, e.g., regulation of autophagy and MAPK signaling networks, 7 out of 10 original gene regulatory networks have major votes in terms of the performance indices by PA estimation, and 9 out of 21 selected defective networks can be corrected based on modification indices. Furthermore, we extend networks which have been correctly identified, and compare the approach with other algorithms such as SSEM algorithm and dynamic Bayesian model method (Luna *et al*., 2007). Compared with SSEM algorithm, the study obtains better results on three networks which are chosen in their study. The true positive rates are 60, 80 and 100%, respectively. Besides, the result is also better than one by dynamic Bayesian model method: the true positive rate is 83.3%. The proposed approach is proved to be able to effectively identify gene regulatory networks, and provide users the degree of confidence on a built network.

REFERENCES

1. 張志榮。2007。以結構方程模型重建基因網路。碩士論文。台北：國立台灣大學醫學工程研究所。

2. Duggan, S. P., W. M. Gallagher, E. J. Fox, M. M. Abdel-Latif, J. V. Revnold, and D. Kelleher. 2006. Low pH induces co-ordinate regulation of gene expression in oesophageal cells. Carcinogenesis. 27(2): 319-327.

3. Forster, J., I. Famili, P. Fu, B. O. Palsson, and J. Nielsen. 2003. Genome-scale reconstruction of the Saccharomyces cerevisiae metabolic network. Genome Res. 13(2): 244-253.

4. Gasch, A. P., P. T. Spellman, C. M. Kao, O. H. Carmel, M. B. Eisem, G. Storz, D. Botstein, and P. O. Brown. 2000. Genomic expression programs in the response of yeast cells to environmental changes. Mol Biol Cell. 11(12): 4241-4257.

5. Gille, C., S. Hoffmann, and H. Holzhuetter. 2007. METANNOGEN: Compiling features of biochemical reactions needed for the reconstruction of metabolic Networks. BMC systems Biology. 1(5).

6. Herrgard, M. J., S. S. Fong, and B. O. Palsson. 2006. Identification of genome-scale metabolic network models using experimentally measured flux profiles. PLOS Comp Biol. **2**: 676-686.

7. Jöreskog, K. G. 1978. Structural analysis of covariance and correlation matrices. Psychometrika. 43: 443-477.

8. Jöreskog, K. G. and D. Sörbom. 2006. LISREL 8.8. Lincolnwood, IL. Scientific Software Inc.

9. Kim, S. Y., S. Imoto, and S. Miyano. 2003. Inferring gene networks from time series microarray data using dynamic Bayesian networks. Briefings In Bioinformatics. 4(3): 228-235.

10. Kanehisa, M., S. Goto, M. Hattori, K. F. Aoki-Kinoshita, M. Itoh, S. Kawashima, T. Katayama, M. Araki, and M. Hirakawa. 2006. From genomics to chemical genomics: new developments in KEGG. Nucleic Acids Res. 34: 354-357.

11. Kim, J., D. G. Bates, I. Postlethwaite, P. Heslop-Harrison, and K. H. Cho. 2007. Least- squares methods for identifying biochemical regulatory networks from noisy measurements. BMC Bioinformatics. 8(8).

12. Luna, I. T., Y. Huang, Y. Yin, D. P. R. Padillo, and M. C. C. Perez. 2007. Uncovering gene regulatory networks from time-series microarray data with variational Bayesian structural expectation maximization. EURASIP Journal on Bioinformatics and System Biology. Article 71312.

13. Murphy, K., and S. Mian. 1999. Modeling gene expression data using dynamic Bayesian networks. Technical Report, University of California at Berkeley, Department of Computer Science.

14. Mehra, S., W. S. Hu, and G. Karypis. 2004. A Boolean algorithm for reconstructing the structure of regulatory networks. Metabolic Engs. 6: 326-339.

15. Novikov, E., and E. Barillot. 2008. Regulatory network reconstruction using an integral additive model with flexible kernel functions. BMC Systems Biology. 2(8).

16. Qiu, Li., S. Leng, Z. Wang, Y. Dai, Y. Zheng, and Z. Wang. 2007. Path analysis of biomarkers of exposure and early biological effects among coke-oven workers exposed to polycyclic aromatic hydrocarbons. Cancer Epidemiol Biomarkers Prev. 16(6).

17. Savageau, M. A. 1969a. Biochemical systems analysis. I. Some mathematical properties of the rate law for the component enzymatic reactions. Journal of Theoretical Biology. 25: 365-369.

18. Savageau, M. A. 1969b. Biochemical systems analysis. II. The steady-state solutions for an n-pool system using a power-law approximation. Journal of Theoretical Biology. 25:370-379.

19. Spellman, P. T., G. Sherlock, M. Q. Zhang, R. I. Vishwanath, K. Anders, M. B. Eisen, P. O. Brown, D. Botstein, and B. Futcher. 1998. Comprehensive identification of cell cycle-regulated genes of the yeast saccharomyces cerevisiae by microarray hybridization. Molecular Cell. 9(12): 3273-3297.

20. Segal, E., M. Shapira, A. Regev, and D. Pe'er, D. Botstein, D. Koller, and N. Friedman. 2003. Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. Nat Genet. 34(2): 166-176.

21. Santos, C. A., D. Senalik, and P. W. Simon. 2005. Path analysis suggests phytoene accumulation is the key step limiting the carotenoid pathway in white carrot roots. Genet Mol Biol. 28: 287-293.

22. Sotiropoulos, V., and Y. Kaznessis. 2007. Synthetic tetracycline-inducible Regulatory networks: computer-aided design of dynamic phenotypes. BMC Systems Biology. 1(7).

23. Shieh, G. S., C. M. Chen, C. Y. Yu, J. Huang, W. F. Wang, and Y. C. Lo. 2008. Inferring transcriptional compensation interactions in yeast via stepwise structure equation modeling. BMC Bioinformations. 9(1): 134.

24. Teixeira, M. C., P. Monteiro, P. jain, S. Tenreiro, A. R. Fernandes, N. P. Mira, M. Alenquer, A. T. Freitas, A. L. Oliveira, and I. S. Correia. 2006. The YEASTRACT database: a tool for the analysis of transcription regulatory associations in Saccharomyces cerevisiae. Nucleic Acids Res. 34: 446-451.

25. Voit, E. O. 2000. Computational Analysis of Biochemical Systems. 1st ed., 37-192. New York: Cambridge University Press.

26. Van Berlo, R., E. P. van Someren, and M. J. T. Reinders. 2003. Studying the conditions for learning dynamic Bayesian networks to discover genetic regulatory networks. Simulation. 79(12): 689-702.

27. Wessels, L., E. P. Van Someren, and M. J. T. Reinders. 2001. A comparison of genetic network models, Biocomputing. 6: 508-519.

28. Wang, J., Y. Huang, M. Sanchez, Y. Wang, and J. M. Zhang. 2006. Reverse engineering yeast gene regulatory networks using graphical models. Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP). 2: 1088-1091.

29. Zou, M. and S. D. Conzen. 2005. A new dynamic Bayesian network (DBN) approach for identifying gene regulatory networks from time course microarray data. Bioinformatics. 21: 71-79.

Appendix-A

Table A-1 Indices for GRN1

		WLS χ^2	AIC	S-RMR	GFI
Original network		35.68	355.68	0.098	0.81
Alternative networks	1	32.83	352.83	0.097	0.81
	2	52.98	372.98	0.092	0.73
	3	65.10	385.10	0.100	0.71
	4	32.13	352.13	0.096	0.81
	5	217.8	537.80	0.150	0.24
	6	36.93	356.93	0.100	0.80
	7	65.10	385.10	0.100	0.71
	8	30.80	350.80	0.098	0.82
	9	52.94	372.94	0.120	0.68
	10	106.3	426.25	0.400	0.57
	11	414.7	734.65	0.580	0.22
	12	96.51	416.51	0.390	0.59
	13	96.08	416.08	0.450	0.59
	14	95.92	415.92	0.420	0.59
	15	96.83	416.83	0.400	0.59
	16	96.45	416.45	0.440	0.59
	17	95.70	415.70	0.420	0.59
	18	97.36	417.36	0.400	0.59
	19	121.4	441.40	0.400	0.39

Table A-2 Indices for GRN2

		WLS χ^2	AIC	S-RMR	GFI
Original network		43.23	363.23	0.015	0.77
Alternative networks	1	55.58	375.58	0.230	0.49
	2**				
	3**				
	4	127.7	447.65	0.110	0.33
	5	266.5	586.48	0.180	0.29
	6	44.60	364.60	0.068	0.76
	7	44.77	364.77	0.069	0.76
	8	116.2	436.15	0.410	0.55
	9	44.60	364.60	0.070	0.76

"**": fail in estimation

Table A-3 Indices for GRN3

		WLS χ^2	AIC	S-RMR	GFI
Original network		128.92	498.92	0.180	0.35
Alternative networks	1	16.01	386.01	0.140	0.91
	2	155.7	525.65	0.170	0.42
	3	16.09	386.09	0.140	0.91
	4	17.02	387.02	0.130	0.90
	5	18.46	388.46	0.100	0.90
	6	39.31	409.31	0.088	0.80
	7	39.38	409.38	0.096	0.80
	8	18.38	388.38	0.088	0.90
	9	44.03	414.03	0.088	0.79
	10	30.23	400.23	0.098	0.84
	11	31.16	401.16	0.120	0.84
	12	30.37	400.37	0.120	0.84
	13	30.18	400.18	0.086	0.84
	14	30.14	400.14	0.087	0.84
	15	31.19	401.19	0.120	0.84
	16	320.4	690.41	0.180	0.26
	17	31.39	401.39	0.086	0.84
	18	59.62	429.62	0.087	0.73
	19	31.36	401.36	0.086	0.84

Table A-4 Indices for GRN4

		WLS χ^2	AIC	S-RMR	GFI
Original network		39.40	359.40	0.012	0.78
Alternative networks	1	66.06	386.06	0.019	0.69
	2	40.30	360.30	0.012	0.78
	3	52.24	372.24	0.011	0.74
	4	43.38	363.38	0.012	0.77
	5	56.96	376.96	0.011	0.72
	6	32.25	352.25	0.012	0.82
	7	54.64	374.64	0.160	0.64
	8	43.32	363.32	0.011	0.77
	9	31.95	351.95	0.011	0.82

Table A-5 Indices for GRN5

		WLS χ^2	AIC	S-RMR	GFI
Original network		21.63	391.63	0.0034	0.88
Alternative networks	1	23.55	393.55	0.0033	0.87
	2	21.66	391.66	0.0034	0.88
	3	21.65	391.65	0.0034	0.88
	4	23.34	393.34	0.0033	0.87
	5	21.71	391.71	0.0034	0.88
	6	23.31	393.31	0.0033	0.87
	7	23.63	393.63	0.0033	0.87
	8	35.27	405.27	0.0055	0.82
	9	68.63	438.63	0.004	0.72

Table A-6 Indices for GRN6

		WLS χ^2	AIC	S-RMR	GFI
Original network		3.10	373.10	0.049	0.98
Alternative networks	1	3.15	373.15	0.049	0.98
	2	3.25	373.25	0.050	0.98
	3	3.16	373.16	0.049	0.98
	4	3.14	373.14	0.049	0.98
	5	3.11	373.11	0.049	0.98
	6	3.76	373.76	0.053	0.98
	7	3.71	373.71	0.053	0.98
	8	3.11	373.11	0.049	0.98
	9	3.71	373.71	0.053	0.98
	10	3.21	373.21	0.049	0.98
	11	3.40	373.40	0.051	0.98
	12	3.39	373.39	0.050	0.98
	13	3.21	373.21	0.049	0.98
	14	3.20	373.20	0.049	0.98
	15	3.39	373.39	0.050	0.98
	16	3.39	373.39	0.050	0.98
	17	3.38	373.38	0.050	0.98
	18	3.72	373.72	0.053	0.98
	19	3.39	373.39	0.050	0.98

Table A-7 Indices for GRN7

		WLS χ^2	AIC	S-RMR	GFI
Original network		6.83	376.83	0.0017	0.96
Alternative networks	1	7.43	377.43	0.0018	0.96
	2	27.6	397.58	0.0180	0.85
	3	6.86	376.86	0.0017	0.96
	4	6.93	376.93	0.0018	0.96
	5	7.64	377.64	0.0017	0.95
	6	25.7	395.71	0.0074	0.86
	7	26.5	396.54	0.0074	0.86
	8	7.24	377.24	0.0018	0.96
	9	25.5	395.54	0.0074	0.86
	10	152	522.45	0.110	0.29
	11	11.9	381.90	0.0018	0.93
	12	11.9	381.91	0.0018	0.93
	13	11.8	381.80	0.0017	0.93
	14	11.7	381.68	0.0018	0.93
	15	11.9	381.91	0.0018	0.93
	16	12.1	382.09	0.0180	0.93
	17	12.2	382.23	0.0017	0.93
	18	33.5	403.48	0.0073	0.83
	19	12.2	382.15	0.0017	0.93

Table A-8 Indices for GRN8

		WLS χ^2	AIC	S-RMR	GFI
Original network		7.950	377.95	0.0050	0.95
Alternative networks	1	49.07	419.07	0.0330	0.75
	2	95.01	465.01	0.2600	0.11
	3	8.090	378.09	0.0050	0.95
	4	8.070	378.07	0.0050	0.95
	5	22.93	392.93	0.0054	0.88
	6	11.53	381.53	0.0051	0.93
	7	11.50	381.50	0.0051	0.93
	8	139.1	509.02	0.0820	0.37
	9	575.3	945.30	0.1400	0.25
	10	23.73	393.73	0.0063	0.87
	11	16.50	386.50	0.0053	0.91
	12	17.61	386.71	0.0053	0.91
	13	16.48	386.48	0.0053	0.91
	14	16.32	386.32	0.0053	0.91
	15	16.53	386.53	0.0053	0.91
	16	16.77	386.77	0.0053	0.91
	17	16.63	386.63	0.0053	0.91
	18	21.25	391.25	0.0053	0.88
	19	16.24	386.24	0.0052	0.91

Table A-9 Indices for GRN9

		WLS χ^2	AIC	S-RMR	GFI
Original network		33.11	403.11	0.0082	0.83
Alternative networks	1	101.6	471.56	0.1000	0.39
	2	140.3	510.27	0.0850	0.39
	3	38.52	408.52	0.0080	0.81
	4	33.77	403.77	0.0081	0.83
	5	33.40	403.40	0.0083	0.83
	6	33.81	403.81	0.0080	0.83
	7	33.81	403.81	0.0080	0.83
	8	136.5	506.52	0.0660	0.41
	9	507.8	877.76	0.2600	0.20

Table A-10 Indices for GRN10

		WLS χ^2	AIC	S-RMR	GFI
Original network		31.01	401.01	0.013	0.84
Alternative networks	1	119.4	489.39	0.110	0.40
	2	31.68	401.68	0.013	0.84
	3	101.9	471.86	0.340	-0.09
	4	31.15	401.15	0.013	0.84
	5	31.54	401.54	0.013	0.84
	6	305.5	675.50	0.070	0.30
	7	31.65	401.65	0.013	0.84
	8	141.4	511.37	0.052	0.35
	9	622.1	992.14	0.290	0.19