國立交通大學

統計學研究所

碩士論文

估計基因網路及基因間函數結構之統計方法

Statistical Methods for Estimation of Genetic Networks and Functional Structures between Genes

研究生：羅巧慧

指導教授：洪志眞　教授

洪慧念　教授

中華民國九十三年六月

估計基因網路及基因間函數結構之統計方法

Statistical Methods for Estimation of Genetic Networks and Functional Structures between Genes

研 究 生：羅巧慧　　Student：Chau-Wui Lo

指導教授：洪志眞 博士　　Advisors：Dr. Jyh-Jen Horng Shiau

洪慧念 博士　　Dr. Hui-Nien Hung

國 立 交 通 大 學

統計學研究所

碩 士 論 文

A Thesis

Submitted to Institute of Statistics

College of Science

National Chiao Tung University

in Partial Fulfillment of the Requirements

for the Degree of

Master

in

Statistics

June 2004

Hsinchu, Taiwan, Republic of China

中華民國九十三年六月

估計基因網路及基因間函數結構之統計方法

研究生：羅巧慧　　指導教授：洪志眞 教授

洪慧念 教授

國立交通大學統計學研究所

摘　要

我們提出一個利用統計方法來建構基因網路的方法。我們將目標分成三個部分。第一部分是先探討兩個變數間是否有關係存在？如果他們之間確實有關係，第二部分就是探討他們是如何的相關？激發或是抑制?接著第三步即是去找出他們相關的方向。在資料分析的部分，利用了蒙地卡羅模擬分析來展示我們所提出方法的分析效率，並且最後我們也把提出的方法應用在酵母菌資料上來作個討論。

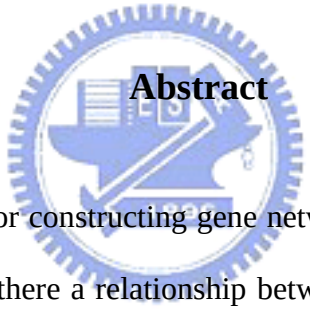
Statistical Methods for Estimation of Genetic Networks and Functional Structures between Genes

Student：Chau-Wui Lo　　Advisors：Dr. Jyh-Jen Horng Shiau

Dr. Hui-Nien Hung

Institute of Statistics

National Chiao Tung University



Abstract

We propose a procedure for constructing gene networks using statistical methods. Three issues are considered. First, is there a relationship between a pair of genes? Second, how do they relate, repress or activate? Third, what is the related direction? By considering the relationship of a pair of genes at a time, our method gives not only the relation (activate, repress) but also the direction for each pair of genes. We conduct Monte Carlo simulations to show the effectiveness of the proposed method. Finally, the method is applied to the Sacharomyces cerevisiae gene expression data as an illustrative example.

誌　謝

在交大這兩年，讓我受益良多，不但學習到專業知識，遇到很多的朋友，更學習到如何解決問題及主動學習的精神。感謝指導教授洪志真老師及洪慧念老師熱心的指導與照顧，也很感謝同學們寶文、崢珮、超毅、欣妤、怡均、淑真、忠庭、政輝、志浩、慶富、文祥、翠英、淑靜、坤民、宏元、宇青的包容和陪伴及博士班學長達叔、宏嘉、牛哥、泰賓的照顧。除此之外很感謝我的家人的支持與鼓勵。能夠順利完成碩士學位，非常感謝各位的支持。

羅 巧 慧　謹誌于

國立交通大學統計學研究所

中華民國九十三年六月

Contents

1. Introduction

In the recent years, a large amount of gene expression data has been collected and estimating a gene network has become one of the popular topics in the field of bioinformatics. The knowledge of the coding sequences of virtually every gene in an organism has enabled the development of technology to simultaneously monitor the expression of all the genes. Due to the curse of dimensionality and complexity of the expression data, it is not an easy task to find structures, which are buried in noise. Several methodologies have been proposed for constructing gene network based on gene expression data, such as Boolean networks [1,2,3,4,18,20], differential equations [6,9], Bayesian networks [10,11,14,15,16,25,26], nonparametric regression [7,8] and a smooth response surface methodology [27]. Some of them will be introduced in reviews. To extract the effective information from micro array gene expression data, theory and methodology are expected to be developed.

We use a simple example to introduce the causal and effect relationship. The height may cause the weight, but the weight cannot influence the height. We can see that there are inspire (active) or depress (repress) between those two variables like the relations between genes. The causal and effect relationship has been discussed in many fields, such as medicine, industry and society. For example, doctors reason the causes of diseases by their accumulated experiences, industrial engineering realize the actual breakdown causes using experimental designs. In this study we are interested in finding a good method to discover the direction and relation between the genes that actually have cause-effect relationship.

We propose a method to find the relationships between genes using statistical methods. We consider the relation of a pair of genes at one time. If there are M genes,

then there are C_2^M times need to be done. Our method gives not only the relations (active, repress) but also the directions between genes. Our method can be divided into two parts by the distributions of the residuals. The second part of our proposed method is to discretize data first. Secondly find whether there is a relation between two variables or not. And also, find how they connected (active, repress). Third, find the directions. Details can be seen from the flowchart latter. The advantages of our method are we don't need strong statistical assumption before using and compute rapidly. Especially when the sample size is small, the resulting graph size is still similar to the graph size of larger sample size. The shortcoming of our method is not sensitive to symmetric functional structures.

The rest of the paper is organized as follows. Section 2 gives a literature review on relate research works. Section 3 decides the proposed methodology. Section 4 presents the results of a simulation study and a real data analysis. Section 5 concludes the paper with a brief summary and discussion.

2. Literature Reviews

2.1 Causal Effect

In this section, we review these methodologies for finding the cause-effect relationships between variables. A simply causal and effect relationship can be decided as follows.

If *X* (cause/parent) causes *Y* (effect/child), then manipulating the value of *X* affects the value of *Y*. On the other hand, if *Y* causes *X*, then manipulating the value of *X* will not affect *Y*. Let $x_i, i = 1,...,n$ be the variables under study. A *functional causal model* in general form consists of a set of equations of the form

$$x_i = f_i(pa_i, u_i), i = 1,....,n \tag{2.1}$$

where pa_i (connoting *parents*) stands for the set of variables judged to be immediate causes of X_i, U_i represents the errors (or "disturbances"), and $f(\cdot)$ is the functional relationship between the variables. Equation (1) is a nonlinear, nonparametric generalization of the linear structural equation models (SEMs).

$$x_i = \sum_{k \neq i} \alpha_{ik} x_k + u_i, i = 1,...,n \tag{2.2}$$

In the linear model, pa_i corresponds to those variables on the right-hand- side of (2) that have nonzero coefficients.

A set of equations in the form of (1) and in which each equation represents an autonomous mechanism is called *structural model*; if each mechanism determines the value of just one distinct variables (called the *dependent* variables), then the model is called a *structural causal model* or a *causal model* for short. To illustrate, Figure 1 depicts a canonical econometric model relating price and demand through the equations

$$q = b_1 p + d_1 i + u_1, \tag{2.3}$$

$$p = b_2 q + d_2 w + u_2, \tag{2.4}$$

where Q is the quantity of household demand for a product A, P is the unit price of the product A, I is the household income, W is the wage rate for producing product A, u_1 and u_2 represent error terms-unmodeled factors that affect the quantity and price, respectively (Goldberger, 1992). The graph associated with this model is cyclic, and the vertices associated with the variables $U_1, U_2, I,$ and W are root nodes, conveying the assumption of mutual independence.

The idea of *autonomy* (Aldrich, 1989), in this context, means that two equations represent two loosely coupled segments of the economy, consumers and producers. Equation (3) describes how consumers decide what quantity Q to buy and (4) describes how manufacturers decide what price P to charge. Like all feedback system, this too represents implicit dynamics; today's prices are determined on the basis of yesterday's demand, and these prices will determine the demand in the next period of transactions. The solution to such equations represents a long-term equilibrium under the assumption that the background quantities, U_1 and U_2, remain constant. [19]

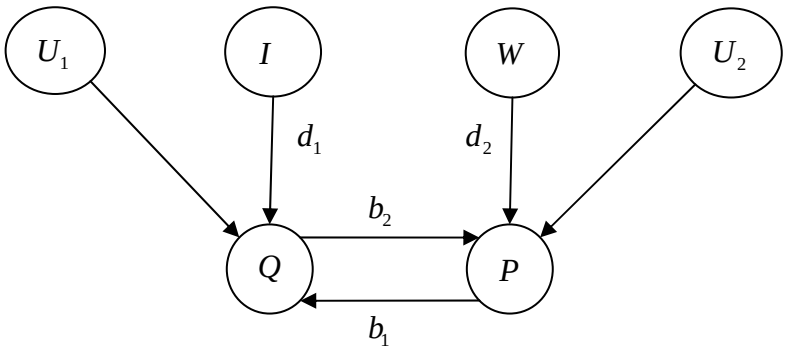


Figure 1: Causal diagram illustrating the relationship between price (P), demand (Q), income (I), and wages (W).

2.2 Bayesian Network

A graph consists a set of *vertices* (or *nodes*) and a set of *edges* (or *links*) that connect some pairs of vertices. The vertices of the graphs correspond to variables and the edges denote a certain relationship that holds in pair of variables. Two variables connected by an edge are celled *adjacent*. If every edge in a path is an arrow that points from the first to the second vertex of the pair, we have a *directed path*. A graph with directed path called is called a *directed graph*. Directed graphs may include directed cycles (e.g., $X \rightarrow Y$, $Y \rightarrow X$), representing mutual causation or feedback process, but no self-loops (e.g., $X \rightarrow X$). A graph that contains no directed cycles is called *acyclic*. A graph that is both directed and acyclic is called a *directed acyclic graph* (DAG). Undirected graphs, sometimes called *Markov networks* (Pearl, 1988b), are used primarily to represent symmetrical spatial relationship (Isham, 1981; Cox and Wermuth, 1996; Lauritzen, 1996). Directed graphs, especially DAGs, have been used to represent causal or temporal relationships (Lauritzen, 1982; Wermuth and Lauritzen, 1983; Kirveri et al., 1984) and are known as Bayesian networks. Figure 2 is an example of a simple Bayesian network structure. This network structure implies several conditional independence statements:

$$I(A;E), I(B;D \mid A,E), I(C;A,D,E \mid B), I(D;B,C,E \mid A), and\ I(E;A,D). \quad (2.5)$$

The network structure also implies that the joint distribution has the product form

$$P(A,B,C,D,E) = P(A)P(B \mid A,E)P(C \mid B)P(D \mid A)P(E). \quad (2.6)$$

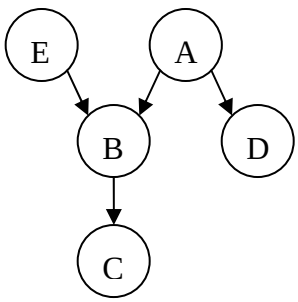


Figure 2: An example of Bayesian network

A causal network can be interpreted as a Bayesian network when we are willing to make the *Causal Markov Assumption*: *given the values of a variable's immediate causes, it is independent of earlier causes.*

2.3 Using Bayesian Network and Nonparametric Regression to Analyze the Causal Effect between Variables

A Bayesian network is a graph-based model of joint multivariate probability distributions that captures properties of conditional independence between variables. It consists two components. The first component, *G*, is a *directed acyclic graph* (DAG) whose vertices correspond to the random variables $X_1,...,X_n$. The second component, θ, describes a conditional distribution for each variable, given its parents in *G*. Together, these two components specify a unique distribution on $X_1,...,X_n$. The graph *G* represents conditional independence assumptions that allow the joint distribution to be decomposed, economizing on the number of parameters. The graph *G* encodes the Markov Assumption: Each variable X_i is independent of its non-descendants, given its parents in *G*.

By applying the chain rule of probabilities and properties of conditional independencies, any joint distribution that satisfies the Markov Assumption can be composed into the *product form*

$$P(X_1,...,X_n)=\prod_{i=1}^{n}P(X_i \mid Pa^G(X_i)), \tag{2.7}$$

where $Pa^G(X_i)$ is the set of the parents of X_i in *G*. The method of Bayesian network is to select a proper graph *G* with the largest posterior probability after giving the prior probability. Imoto et al. (2002) proposed the criterion

$$BNRC(G)=-2\log\left\{\pi_G\int\prod_{i=1}^{P}f(x_i \mid \theta_G)\pi(\theta_G \mid \lambda)d\theta_G\right\} \tag{2.8}$$

where π_G is the prior probability of the graph *G*, $\pi(\theta_G \mid \lambda)$ is the prior probability of

θ_G after giving parametric vector λ, $\theta_G = (\theta_1^T, ..., \theta_n^T)$ is the parametric vector in *G*, θ_j is the parametric vector of model f_j, and

$$f(x_i \mid \theta_G) = \prod_{j=1}^{P} f_j(x_{ij} \mid p_{ij}, \theta_j) \tag{2.9}$$

$$f_j(x_{ij} \mid p_{ij}, \gamma_j, \sigma_j^2) = \frac{1}{\sqrt{2\pi\sigma_j^2}} \exp\left[-\frac{\left(x_{ij} - \sum_{k=1}^{q_j} \sum_{m=1}^{M_{ij}} \gamma_{mk}^{(j)} b_{mk}^{(j)}(p_{ik}^{(j)}) \right)^2}{2\sigma_j^2} \right] \tag{2.10}$$

where x_{ij} is the *i*th observation of *j*th variable, that is, if the number of observation is *n* then the observation of the *j*th variable X_j is $x_{1j}, \ldots, x_{nj}$. We use $P_{1j}, ..., P_{nj}$ to represent the parents of X_j. p_{ij} is the vector with q_j-dimension, i.e., the *j*th variable has q_j parents at the *i*th observation, and its *k*th component is $p_{ij}^{(k)}$. And use nonparametric regression models for capturing the relationship between x_{ij} and $p_{ij} = (p_{i1}^{(j)}, ..., p_{iq_j}^{(j)})$ in the form

$$x_{ij} = m_1(p_{i1}^{(j)}) + m_2(p_{i2}^{(j)}) + ... + m_{q_j}(p_{iq_j}^{(j)}) + \varepsilon_{ij}, i = 1, ..., n; j = 1, ..., p. \tag{2.11}$$

where $m_k (k = 1, ..., q_j)$ are smooth functions from R to R, and $\varepsilon_{ij} (i = i, ..., n)$ depend independently and normally on mean 0 and variance σ_j^2. For m_k, it is assumed that

$$m_k(p_{ik}^{(j)}) = \sum_{m=1}^{M_{jk}} \gamma_{mk}^{(j)} b_{mk}^{(j)}(p_{ik}^{(j)}), i = 1, ..., n; k = 1, ..., q_j. \tag{2.12}$$

where $\left\{ b_{1k}^{(j)}, ..., b_{M_{jk}k}^{(j)} \right\}$ is a prescribed set of basis functions (such as Fourior series, polynomial bases, regression spline bases, B-spline bases, wavelet bases and so on), $(\gamma_{1k}^{(j)}, ..., \gamma_{M_{jk}k}^{(j)})$ is the unknown coefficients and M_{jk} is the number of basis. Bayesian networks are to find *G* makes *BNRC* maximized.

Recently, Imoto et al. (2002) porposed the use of nonparametric additive regression

models for capturing not only linear dependencies but also nonlinear structures between genes. Imoto et al. (2003) improved this method by using Bayesian networks and the nonparametric heteroscedastic regression, which is more resistant to the effect of outliers but it needs much time for determining the optimal graph. Tamada et al. (2003) proposed a motif detection method to estimate gene networks and combine microarray gene expression data and DNA sequences of regulatory regions of genes. It found that the motif information is useful for revising some incorrect relations in the network estimated by microarray alone. Imoto et al. (2003) proposed a method for estimating a gene network based on Bayesian networks from microarray gene expression data together with biological knowledge including protein-protein interactions, protein-DNA interactions, binding site information, existing literature and so on. Its proposed criterion can control the trade-off between microarray information and biological knowledge automatically. Kim et al. (2003) proposed a Bayesian network and nonparametric regression model for constructing a gene network from time series microarray gene data. This method can overcome a shortcoming of the Bayesian network model in the sense of the construction of cyclic regulations.

2.4 Using Nonparametric Regression Method to Analyze the Causal Effect between Variables

Different from linear regression analysis, nonparametric regression uses a roughness penalty that decreases as the fitting curve gets smoother.

Consider the *n* observations $\{(x_i, y_i), i = 1,...,n\}$. Assume that the sample is ordered over the interval [*a, b*] with respect to the predictor values; that is, $a \le x_1 \le ..... \le x_n \le b$. To estimate the unknown smooth regression function by explicitly trading off fidelity to the data with smoothness of the estimate. For regression, the residual sum of squares is a natural measure of fidelity to the data, so the roughness

penalty estimator is the minimizer of

$$S(g) = \frac{1}{n}\sum_{i=1}^{n}(y_i - g(x_i))^2 + \lambda\int_a^b (g''(x))^2 dx, \tag{2.13}$$

and the resulting cubic smoothing spline estimate $\hat{g}$ belongs to the Sobolev space

$$W_2^{\,2}[a,b] = \{g \mid g \text{ and } g' \text{ are absolutely continuous, g'' is square integrable.}\} \tag{2.14}$$

Where $\lambda > 0$ is the smoothing parameter. If $\lambda = 0$, the smoothing spline becomes an interpolating spline that passes through each of the responses y_i, while if $\lambda \to \infty$, $\hat{g}$ approaches the linear least squares regression line. The smoothing spline is a linear estimator, so the vector of fitted values $\hat{y}_i = \hat{g}(x_i)$ can be written as $\hat{y} = A(\lambda)y$. The matrix $A(\lambda)$ is called the hat matrix. λ can be chosen to minimize the *cross-validation score* [17]

$$CV(\lambda) = \frac{1}{n}\sum_{i=1}^{n}[y_i - \hat{g}^{(i)}(x_i)]^2, \tag{2.15}$$

where $g^{(i)}(x_i)$ is the spline estimate based on all the observations except x_i, evaluated at x_i. It can be shown that for linear smoothers, $CV(\lambda)$ can be written as a function of the fitted values (Green and Silverman, 1995) [17],

$$CV(\lambda) = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{y_i - \hat{g}(x_i)}{1 - A_{ii}(\lambda)}\right]^2, \tag{2.16}$$

where $A_{ii}(\lambda)$ is the ith diagonal element of the hat matrix $A(x)$. $A_{ii}(\lambda)$ is called the leverage value at x_i since it measures the potential for y_i the observed response to exert influence on the fitted value.

A variation is *generalized cross-validation (GCV)* [17], which replaces each value $1 - A_{ii}(\lambda)$ with their average, $1 - n^{-1}trace[A(\lambda)]$. The generalized cross-validation selector of λ, $\hat{\lambda}_{GCV}$, is the minimizer of

$$GCV(\lambda) = \frac{\sum_{i=1}^{n}[y_i - \hat{g}(x_i)]^2}{n\{1 - n^{-1}trace[A(\lambda)]\}^2}. \tag{2.17}$$

Proposition 1. (Shiau, 1985) Let $\hat{g}_\lambda$ be the smoothing spline estimator. Then

$$S(\hat{g}_\lambda) = y^T (I - A(\lambda)) y / n, \tag{2.18}$$

The following two propositions are given in Cheng (2003).

Proposition 2. $\hat{g}_\lambda = \arg_g \min S(g)$, and is the mode of the posterior density after giving the prior distribution.

Proposition 3. Assume that X and Y and *E(X)=E(Y)=0, Var(X)=Var(Y)=1* , let $(x_i, y_i), i = 1,...,n$, be *n* i.i.d. observations. Then $E(S(\hat{g}_\lambda)) \to 1, as\ n \to \infty.$

Cheng (2003) consider only the causal relationship between two variables X and Y. If X is the parent (cause) and Y is the child (effect), then denote the cause-effect relationship by X→Y. more specifically, the causal relationship between X and Y can be described by the following causal models: $Y = g_1(X) + \varepsilon$, where g_1 is a smoothing function and ε is a random error with mean zero and independent of X. If X cannot be simultaneously represented as $g_2(Y) + \varepsilon'$ with Y independent of an random error ε', then X→Y but not vice versa.

We shall call the method proposed in Cheng (2003) determing the causal direction between two random variables “SCORE” method. SCORE method simultaneously use nonparametric method to estimate g_1 and g_2 . Finally gets two score $S(\hat{g}_1)$ and $S(\hat{g}_2)$ (Shiau, 1985), then use *decision rules* to determine the direction.

We simplify Chia Yu Cheng et al. (2003) as the following flow chart.

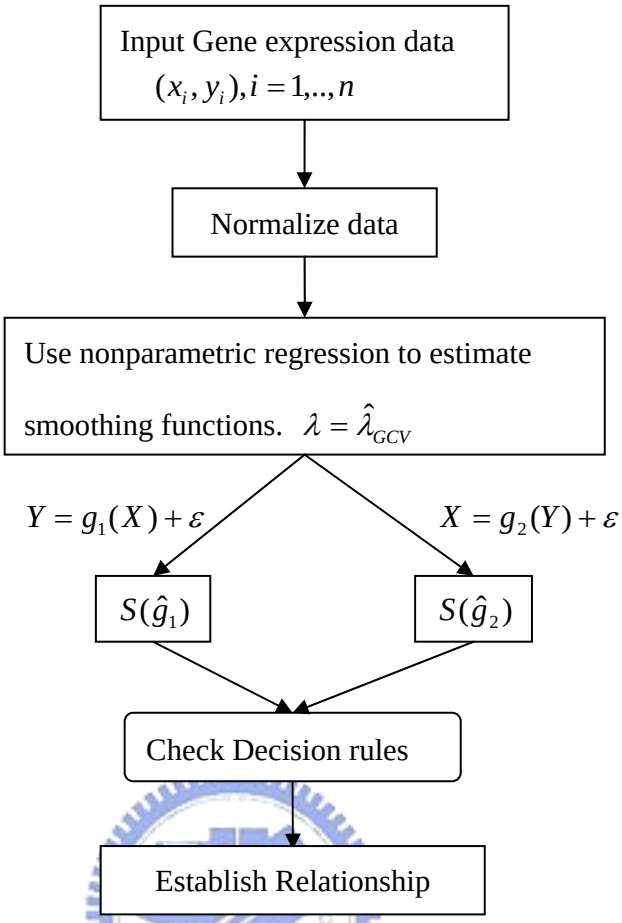


Figure 3: The flowchart of SCORE

Decision rules are as follows*:*

If $S(\hat{g}_1) < S(\hat{g}_2)$ then $X \rightarrow Y$ has a larger posterior probability than $Y \rightarrow X$, i.e., X causes Y.

If $S(\hat{g}_1) > S(\hat{g}_2)$ then $Y \rightarrow X$ has a larger posterior probability than $X \rightarrow Y$, i.e., Y causes X.

If $S(\hat{g}_\lambda)$ is close to 1, then $X \perp Y$ (Cheng (2003) suggested $S(\hat{g}_\lambda) > 0.9$, and Lu (2003) suggested $S(\hat{g}_\lambda) > 0.8$).

We call the 0.9 or 0.8 “the threshold score of independence”.

2.5 Using Smooth Response Surface Methodology to Construct Gene Networks

A smooth response surface algorithm proposed by Xu et al. (2002) is a sophisticated data mining technique for analyzing gene expression data and constructing gene networks. It uses a three-dimensional smooth response surface to capture the biological relationship between the target and activator-repressor. It also functionally

describes triplets of activators, repressors, and targets, and their regulations in gene expression data. The diagnostic strategy in the algorithm is to evaluate the scores of the triplets so that those with low scores are kept and a regulatory network is constructed based on this information and existing biological knowledge. Xu et al. (2002) applied the method to two yeast gene expression data sets and reported that the predictions based on the identified triplets agree with some experimental data in the literature. This method provides a novel model with attractive mathematical and statistical features that make the algorithm valuable for mining expression or concentration information, help for determining the function of uncharacterized proteins, and can result in a better understanding of coherent pathways.

The smooth response surface method is decided as follows. First, transform the raw data over the time points into the interval [*0, 1*]. Let A be the activator and B be the repressor. The definition of the activator-repressor-target model is that a target gene C is controlled by both an activator gene A and a repressor gene B. Define a three-dimensional smooth response surface as *S(A, B)*, which is a piecewise linear-quadratic polynomial on [*0, 1*] × [*0, 1*]. The triplets that follow the activator-repressor-target relationship should lie closely to the response surface.

The 3D response surface has the same purpose with a high dimensional decision matrix. The function *S(A, B)* maps two normalized values *A* and *B* onto a 3D surface, in order to describe a surface response value *C* .To ease decision-making, it uses some heuristic rules:

A high + *B* low → *C* high.

A low + *B* high → *C* low.

As seen in Figure 3, a triplet *(A, B, S(A, B))* represents the biological relationship that follows the pattern of a target *S(A, B)* controlled by an activator *A* and a repressor *B* as described in the activator-repressor-target model. The response surface captures the

biological model with features such as compactness, simplicity and visualization. The imputation and gene filtering steps are applied to remove noise from the data.

For each triplet *(A, B, C)*, $\hat{C} = S(A,B)$ is the fitted value of a target *C*. The residual, $\hat{C} - C$, should be small, if the activator-repressor-target relationship is strong. The residual sum of squares measures the overall variation in *C* that is not explained in the response surface model. Then the lack-of-fit function *RT(A, B, C)*, i.e., the ratio of the residual sum of squares and the total sum of squares, describes the proportion of variation in *C* that is not captured by the 3D response surface. A small value of lack-of-fit indicates that there is a strong activator-repressor-target relationship among *A*, *B*, and *C*. The lack-of-fit formula is defined to filter the triplets in the initial screening. To save storage and computation, only those triplets whose lack-of-fit values do not exceed a given constant *RT* are kept.

A diagnostic strategy *Diag(A, B, C)* is applied to check the reliability of the triplets after the initial filtering to measure robustness of the fitted model for each triplet *(A, B, C)*. Xu et al. (2002) pointed out that the intensity measurement of gene expression at one or two time points may deviate from the model and suggest that the measurement may be faulty and should be treated as an outlier. If such a value occurs at the *i*-th point, then $RT_{(i)}(A,B,C)$, i.e., the lack-of-fit of *(A, B, C)* when the *i*-th point (or the *i*-th column) is left out, will differ greatly from *RT(A, B, C)*. *Diag(A, B, C)* provides a summary measure over all time points for a given triplet. The diagnostic method is developed to refine the selected triplets and a score, which reflects the strength of the triplet interrelationship, is defined to rank the refined triplets. A larger *Diag* value would suggest that the information for the triplet is unreliable and should be removed for further consideration. Thus the criteria for selecting triplet candidates is: $RT(A,B,C) \le RT$ and $Diag(A,B,C) \le Diag$, where *RT* and *Diag* are constants as

specified by users.

A final score is defined to measure the strength of the triplet interrelationship. *Score(A, B, C)* is a function of the lack-of-fit value and the diagnostic measure, and focuses primarily on the *RT(A, B, C)* value and secondly on the *Diag(A, B, C)* values. Triplets with low values of *RT(A, B, C)* and *Diag(A, B, C)* will have low scores, which indicate a close relationship among *A*, *B*, and *C*. Finally, a gene regulatory network is constructed based on the top scoring triplets. Figure 4 gives the flowchart of the smooth response surface algorithm.

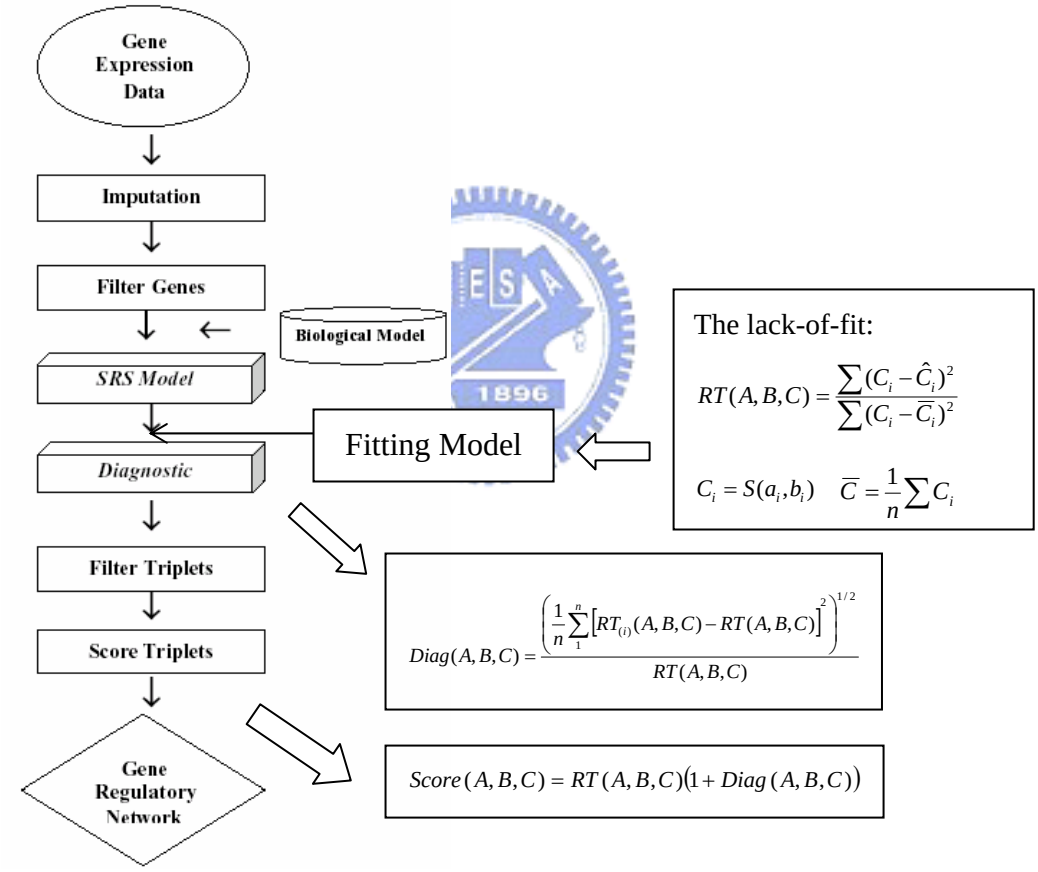


Figure 4: The data processing flowchart of the *Smooth Response Surface* algorithm.

3. Methodology

3.1 Motivation

Xu et al. (2002) proposed the activator-repressor-target model to represent that each variable can find its activator and repressor and is controlled by the activator and repressor simultaneously. It is not clear whether there exists such a triplet relationship in reality. In this study, we consider pairs of genes instead of triplets.

Cheng (2003) proposed a score method to determine the relationship between two variables. But this method only determines the causal effect direction but no activate (+) / repress (-) information. Also, there is a problem of determining whether two variables are independent. Therefore, we use the method in next section to improve Cheng's method.

3.2 Proposed Methods

We simplify our goal as finding the relation between two variables.

If X is the parent (cause) and Y is the child (effect), then we use X→Y to represent the causal effect relation between them. We assume X and Y have the regression relation: $Y = f(X) + \varepsilon$, where f is a smoothing function and ε is a random variable with mean zero and is independent of X. Sometimes, X can be represented as $g(Y) + \varepsilon'$ with Y being independent of ε' at the same time. This will hold when f is a linear function. When this happens, we denote the cause/effect reality by $X \leftrightarrow Y$.

Let R_1 and R_2 be the residuals of regressory Y on X and X on Y, respectively. The core idea of our method is that, if X→Y, then we should have the following result:

$$R_1 \perp X \text{ and } R_2 \not\perp Y, \text{ where } R_1 = Y - \hat{f}(X) \text{ and } R_2 = X - \hat{g}(Y).$$

On the other hand, if Y→X, then $R_1 \not\perp X$ and $R_2 \perp Y$.

The possible relationships between X and Y are:

$X \xrightarrow{+} Y,\ X \xrightarrow{-} Y,\ X \xleftarrow{+} Y,\ X \xleftarrow{-} Y,\ X \xleftrightarrow{+} Y,\ X \xleftrightarrow{-} Y,\ X \perp Y.$

Arrow is the direction. The sign above the arrow represents activate (+) or repress (-).

Our proposed method is divided into *two parts* by the patterns of the residuals R_i, i=1, 2.

We summarize the method by the following flow chart.

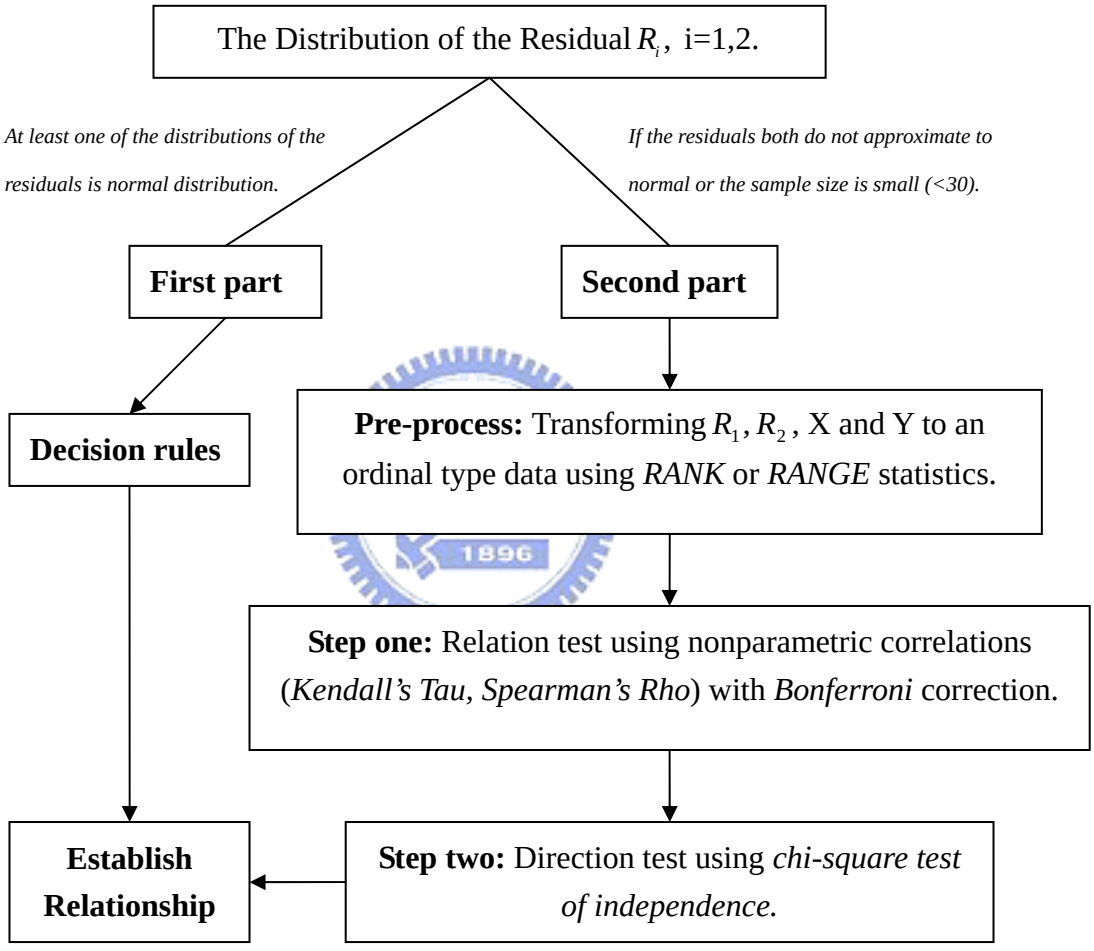


Figure 5: The flow chart of our proposed method

We can use a graphical display, *QQ plot,* to check the normality of the residuals. *QQ plots* are used to assess whether a data set has a particular distribution, or whether two datasets have the same distribution. If two distributions are the same, then the plot will approximate a straight line. The extreme points have more variability than points toward the center. Or we can use *Kolmogorov-Smirnov Goodness-of-Fit Test*

(Chakravart, Laha, and Roy, 1967) [5] and *Chi square Goodness-of-Fit Test* (Snedecor and Cochran, 1989) [24] to compare the distribution of the residuals with normal. If the sample size is small, we suggest skip this part and go to second part.

First part: *At least one of the distributions of the residuals is normal distribution.*

(We remark that this condition usually does not hold in our study.)

In this part we use the following Decision rule:

1. If R_1 is approximately normal and R_2 isn't, then $X \rightarrow Y$.
2. If R_2 is approximately normal and R_1 isn't, then $Y \rightarrow X$.
3. If R_1 and R_2 both are approximately normal, then $X \leftrightarrow Y$.
4. If R_1 and R_2 both are not approximately normal or the sample size is small, then go to second part.

From our experience, the chance of using the first part of our method is quite small.

Second part: *When the residuals* R_1 and R_2 *both are not approximately normal or the sample size is small (<30).*

In this part, we divide the target into the following steps:

Q1. Is there a relationship between those two variables?

Q2. How do they relate? Repress (-) or activate (+)?

Q3. What is the related direction if they really have relationship?

Step one will solve the Q1 and Q2. Step two will give the answer to Q3.

If we want to know the relationship between two variables, we must confirm whether there is relation between two variables first.

Relationships between variables. To express a relationship between two variables, one way is to compute the correlation coefficient between two variables. We discuss the correlation between *X* and *Y* using nonparametric correlations (*Kendall Tau and Spearman R*). An advantage of nonparametric or rank correlation is that we need not

know the probability distribution functions from which the x_i's and y_i's are drawn. However, the slight loss of information in ranking is a small price to pay for a very major advantage: when a correlation is demonstrated to be present nonparametrically, then it is really there! Nonparametric correlation is more robust than linear correlation, more resistant to unplanned defects in the data

Spearman Rank-Order Correlation Coefficient (Siegel & Castellan, 1988 and Siegel, 1956) [23, 24]

Suppose we have *N* data points (x_i, y_i), $i = 1, ..., N$. Let R_i be the rank of x_i among the other *x*'s, S_i be the rank of y_i among the other *y*'s, then the rank-order correlation coefficient is defined to be the linear correlation coefficient of the ranks, namely,

$$r_s = \frac{\sum_i (R_i - \bar{R})(S_i - \bar{S})}{\sqrt{\sum_i (R_i - \bar{R})^2} \sqrt{\sum_i (S_i - \bar{S})^2}} \tag{3.1}$$

If *N* is larger than 10, the significance of a nonzero value of r_s is tested by computing

$$t = r_s \sqrt{\frac{N-2}{1-r_s^2}} . \tag{3.2}$$

This statistic is distributed approximately as Student's t distribution with *N -2* degrees of freedom. A key point is that this approximation does not depend on the original distribution of the x_i's and y_i's; it is always the same approximation, and always pretty good.

Kendall's Tau (Helsel & Hirsch, 1995) [12]

Suppose we have *N* data points (x_i, y_i). Now consider all $\frac{1}{2}N(N-1)$ pairs of data points, where a data point cannot be paired with itself, and where the points in either order count as one pair. Let (x_i, y_i) and (x_j, y_j) be a pair of (bivariate) observations. If $x_i - x_j$ and $y_i - y_j$ have the same sign, we say that pair is *concordant*. If they have opposite signs, we say that the pair is *discordant.*

Let *C* be the number of *concordant* pairs, and *D* be the number of *discordant* pairs, then Kendall's Tau is defined as

$$\tau = \frac{C - D}{\binom{N}{2}}. \tag{3.3}$$

If $x_i = x_j$, or $y_i = y_j$, or both, the comparison is called a *'tie'*. Ties are *not* counted as *concordant* or *discordant*. If the number of ties is large, then Tau has to be replaced by

$$\tau = \frac{C - D}{\sqrt{\left[\binom{N}{2} - N_x\right] \times \left[\binom{N}{2} - N_y\right]}}, \tag{3.4}$$

where N_x be the number of ties involving *x* and N_y be the number of ties involving *y*. Obviously, $-1 \le \tau \le 1$.

If *N* is larger than 40, the significance of *Kendall's Tau* can be tested by calculating a test statistic, *t*, and compares it to the tabular values of *Student's t distribution*:

$$t = \frac{\tau}{\sqrt{\frac{2 \times (2 \times N + 5)}{9 \times N \times (N - 1)}}}. \tag{3.5}$$

We use $H_a : \tau \neq 0$ as the alternative hypothesis at the step one of the second part of our proposed method.

Kendall's Tau is equivalent to *Spearman's Rho* (3.1) with regard to the underlying assumptions. But they are not equal in magnitude because their underlying logic and computational formulae are quite different. They have a relation represented as

$$-1 \le 3 \times \tau - 2 \times r_s \le 1 \tag{3.6}$$

In order to use the rank correlation test, we must transform the original data to ordinal type data. The following pre-process is necessary.

Pre-process: Transforming R_1, R_2, X, and Y to ordinal type. We use the following methods:

1. RANK:

Let $X_1,...,X_N$ be continuous data. R_i is the rank of X_i. We use $Int\left[\frac{R_i}{N/K}\right]$ to translate the original continuous data to a discrete ordinal data type with *K* classes.

2. RANGE:

Let $X_1....X_N$ be continuous data. *A*= max ($X_1....X_N$), *B*=min ($X_1....X_N$). We use $Int\left[\frac{X_i - B}{A-B}\times(1-10^{-7})+5\times10^{-8}\right]\times K$ to translate the original continuous data to discrete ordinal type data with *K* classes.

Step one: Relation test.

We use those two nonparametric correlations (*Kendall's Tau* and *Spearman's Rho)* to perform the rank correlation test. The null hypothesis is that the coefficient (*Kendall's Tau* or *Spearman's Rho)* is zero. We use the signs of those two coefficients to indicate the way of connection (repress (-) or activate (+)). Next we use *Bonferroni correction* to combine the above results.

The rank correlation test is a distribution-free test that determines whether there is a monotonic relation between two variables. A monotonic relation exists when any increase in one variable is invariably associated with either an increase or a decrease in the other variables.

Bonferroni Correction (Sidak, 1968, 1971) [21, 22]

The following is the *Bonferroni general inequality:*

$$P(\bigcap_{i=1}^{g} A_i) \geq 1 - \sum_{i=1}^{g} P[\overline{A}_i], \tag{3.7}$$

where A_i and its complement $\overline{A}_i$ are any events, g is the number of statements or comparisons in the finite set. In particular, if each A_i is the event that a calculated confidence interval for a particular linear combination of treatments includes the true value of that combination, then the left-hand side of the inequality is the probability that all the confidence intervals simultaneously cover their respective true values. The right-hand side is one minus the sum of the probabilities of each of the intervals missing their true values. Therefore, if simultaneous multiple interval estimates are desired with an overall confidence coefficient 1-α, one can construct each interval with confidence coefficient (1-α/g), and the Bonferroni inequality ensures that the overall confidence coefficient is at least 1-α.

In our simulations, we use 0.0975 for α. So if we apply a significance level of 0.05 to each of the two tests, there is now only a 5% chance that any of them will be declared significant under the null hypothesis.

If the Step one rejects the null hypothesis, then we can say that the relations are not strong enough to be noted and the two variables are uncorrelated. Otherwise, we must go a step further to differentiate what the related direction is.

Step two: Direction test.

At this step, we only focus on which direction is better, $X \rightarrow Y$, $Y \rightarrow X$, or $X \leftrightarrow Y$. We decide the direction by comparing the strength of independence of R_1 and X with that of R_2 and Y. We use *Pearson chi-square test* to examine the independence of two

variables.

We can intuitively know that if the smooth regression function fits the data well, then the residual should be small and is (almost) independent of the predictor variable. Use *Pearson chi-square test* to examine the dependence of two variables. Note that the real value of p-value will miss its essential meaning under some conditions. Its validity depends heavily on the assumption that the expected cell counts are at least moderately large; a minimum size of five is often quoted as a rule of thumb. Even when cell counts are adequate, the chi-square is only a large-sample approximation to the true distribution of X-squared under the null hypothesis. We only need to compare the magnitudes of the two p-values, or we can use the Pearson's X-squared statistic directly. So, even if the sample size is small, we still can use this method. The discriminant rules are as follows:

Let P_1 and χ^2_1 be the P-value and X-squared statistic of the *Pearson's chi-square test* of R_1 and X, P_2 and χ^2_2 be the P-value and X-squared statistic of the *Pearson's chi-square test* of R_2 and Y, respectively.

1. If P_1 is larger than P_2, then and we accept $X \to Y$. That means R_2 &Y less independence than R_1 &X. (i.e. If χ^2_1 is smaller than χ^2_2, than we accept $X \to Y$.)
2. If P_2 is larger than P_1, then we accept $Y \to X$. That means R_1 &X less independence than R_2 &Y. (i.e. If χ^2_1 is larger than χ^2_2, then we accept $Y \to X$.)
3. If $|P_1 - P_2| < 0.00001$, then we accept $X \leftrightarrow Y$. The dependence of R_1 &X is similar to that of R_2 &Y. (i.e., If χ^2_1 is very close to χ^2_2, then we accept $X \leftrightarrow Y$.)

4. Empirical Studies

4.1 Monte Carlo Simulation

1. Data generation: (Imoto et al, 2002)

$$X_1 = X_2^{\,2} + 2\sin(X_5) - 2X_7 + \varepsilon_1,\ X_2 = \frac{1}{1+\exp(-4X_3)} + \varepsilon_2$$

$$X_3 = \varepsilon_3,\ X_6 = \varepsilon_6,\ X_9 = \varepsilon_9,\ X_4 = \frac{X_5^{\,2}}{3} + \varepsilon_4,\ X_5 = X_3 - X_6^{\,2} + \varepsilon_5$$

$$X_7 = \begin{cases} -1+\varepsilon_7, & \mathrm{X}_8 \le -0.5 \\ X_8+\varepsilon_7, & -0.5 < \mathrm{X}_8 \le 0.5 \\ 1+\varepsilon_7, & 0.5 < \mathrm{X}_8 \end{cases}$$

$$X_8 = \frac{\exp(-X_4 - 1)}{2} + \varepsilon_8,\ X_{10} = \cos(X_9) + \varepsilon_{10}$$

$$\varepsilon_i \overset{i.i.d.}{\sim} N(0,1),\ \mathrm{i} = 1....10$$

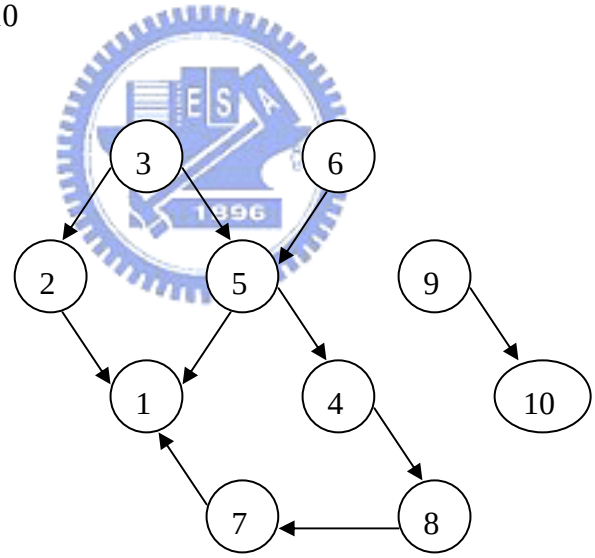


Figure 6: True graph

2. Analysis Methods:

Recall that the Cheng's method (2003) only can find the causal effect direction but not the connected relation (repress or activate). Also, we discovered the threshold score of independence needs to be adjusted for different sample sizes. If we still use 0.9 as the threshold score, the graph size will be much larger than Figure 6 and there are too many

extra pathways. Also, the size of the resulting graph will be very large, and the Cheng's method will lose its practicability. When the sample size is less than or equal to 100, we choose the threshold score of independence by considering the graph size of Figure 6. Therefore, the results of SCORE with the sample size less than 100 are adjusted. Table 6 is the threshold score of independence discovered in our simulation. (The threshold score will be used later in the real data analysis).

We modify Cheng's method as described in the following:

(1) Score evaluation is the same, $S(\hat{g}_{\lambda})$ in (2.17).

(2) The threshold score (Table 6) changes with the sample size.

(3) Using the signs of the nonparametric correlation coefficients (*Kendall's Tau*, *Spearman's Rho*) to indicate the way of connection (repress (-) or activate (+)).

(4) If one of the smoothing parameters is smaller than 10^{-6}, we replace it by $\lambda_{new} = \frac{(\lambda_1 + \lambda_2)}{2}$, where λ_1 and λ_2 are the smoothing parameters used in X→Y and Y→X, respectively.

We use "RANK" and "RANGE" to represent our methods with different pre-processing. Using "SCORE" represents the method modified Cheng's method. We compare our proposed method (RANK and RANGE) with the modified Cheng's method (SCORE).

We use different numbers of sample sizes (*N*) and the classes of ordinal type data (*K*) to see the performance of our methods and Cheng's method.

Cheng (2003) has mentioned that if lambda (the smoothing parameter) is too small, then the fitting curve will be too rough. If lambda is too small, the influence of penalty term will be small. The curve will completely go with the data. This may cause the curve undersmooth. She also suggests that if one of the $\lambda's$ is smaller than 10^{-6}, it should be adjusted by a new lambda, $\lambda_{new} = \frac{(\lambda_1 + \lambda_2)}{2}$ or $\lambda_{new} = \sqrt{\lambda_1 \lambda_2}$ where λ_1 and λ_2

are the smoothing parameters used in X→Y and Y→X, repectively. We take her idea and use $\lambda_{new} = \frac{(\lambda_1 + \lambda_2)}{2}$ as our new lambda when the sample size is less than 100 and lambda is smaller then 10^{-6}.

3. Simulation Results

(1) When the sample size (*N*) is less than 100, the graph size of SCORE is smaller than RANK and RANGE. The threshold values of independence of SCORE are given in Table 4. (Figure 28~Figure 34)

(2) When *N* is larger than 100, the graph size of SCORE is larger than RANK and RANGE. The threshold limits of independence of SCORE we used are 0.9. (Figure 22~Figure 27)

(3) If we check the functions that generate the data, we can see that there are still some atavistic relationships between those variables. Figure 6 does not show this relation. RANK, RANGE, and SCORE can show some of these atavistic relations. We find that RANK and RANGE is more sensitive than SCORE with this kind of relation when the sample size is larger than 100.

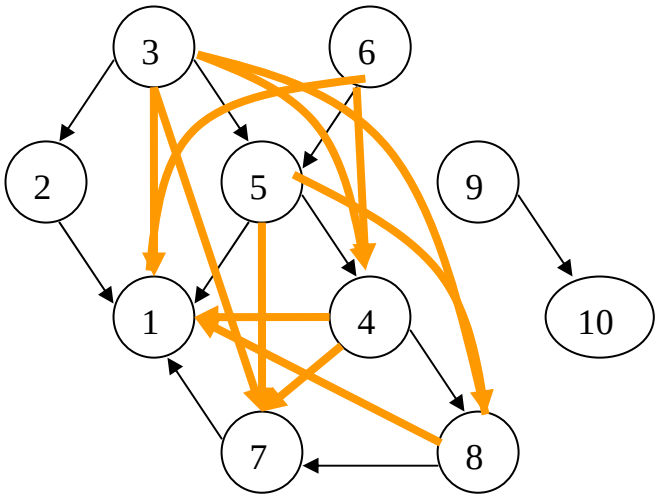


Figure 7: Figure 6 with some atavistic relationships

(4) The graph size of RANK and RANGE decreases as the sample size decreases. (Figure 25 (D)→Figure 28 (C)→Figure 29 (C)→Figure 31 (C)) The graph size

of SCORE increases as the sample size decreases. When the sample size is less than 100, the graph size of SCORE grows very quickly as *N* decreases. (Figure 22 (A) → Figure 12 (A) → Figure 24 (A) → Figure 25 (A) → Figure 27 (A)→ Figure 28 (B) → Figure 29 (B) → Figure 31 (B) → Figure 33 (B))

(5) Check the results of RANGE and RANK. The number of wrong directions when $N \le 100$ (Table 4) is *larger* than that when $N \ge 100$ (Table 2). That is, the number of wrong directions increases when the sample size decreases.

(6) For both RANK and RANGE methods, when $N \le 100$, the number of extra pathways (Table 4) is *smaller* than that when $N \ge 100$ (Table 2). That is, the number of extra pathways decreases with the sample size decreases. The result of SCORE is opposite.

(7) For RANK and RANGE without adjusting lambda, the number of missing pathways when $N < 100$ (Table 4) is not large.

(8) We find that it is not necessary to adjust lambda with RANK and RANGE. But it is necessary to adjust lambda with SCORE, especially when the sample size is less than 100. (Compare Figure 33 (A) with Figure 33 (B))

(9) We find that there are still some relations not detected by RANK and RANGE even if the sample size is large. The reason is that the first-step correlation test cannot detect some symmetrical functions. (For example, $f(x) = x^2$. Details can be seen in Section 4.2.) Although we may not be able to detect those relationships in step one, we can adjust the result by the second step. If the functional structure between two variables is symmetrical and we cannot detect it from the first step, the p-value of chi-square test of one direction in the second step will be very close to zero, (e.g., <0.0001). We give the following example: Table 1 (partial result of Table 10) shows that the p-values of the first step are larger than 0.0975. Our method indicates that both (5, 6) and (9, 10)

have no strong relation. But Figure 8 shows that there is a relation between (5, 6) (Figure 8 (C)) and between (9, 10) (Figure 8 (E)). At the second step, the p-value of the chi-square test is *about zero* in *one* of the directions. We find that this condition indicates that some relations are missed and we must regulate the results of the step-one and change the conclusion to $6 \xrightarrow{+} 5$ and $9 \xrightarrow{+} 10$.

N=200, K=8. (From Table 2)

Table 1: The partial results of the Monte Carlo simulation of N=200 with K=8.

Procedure			(5,6)	(9,10)
First step		Bonferroni correction (p-value=0.0975)	**0.4072**	**0.9984**
		sign	+	+
Second step	RANGE	The p-value of chi-square test of R_1 & X	**<0.0001**	0.34
		The p-value of chi-square test of R_2 & Y	0.85	**<0.0001**
	RANK	The p-value of chi-square test of R_1 & X	**<0.0001**	0.81
		The p-value of chi-square test of R_2 & Y	0.23	**<0.0001**

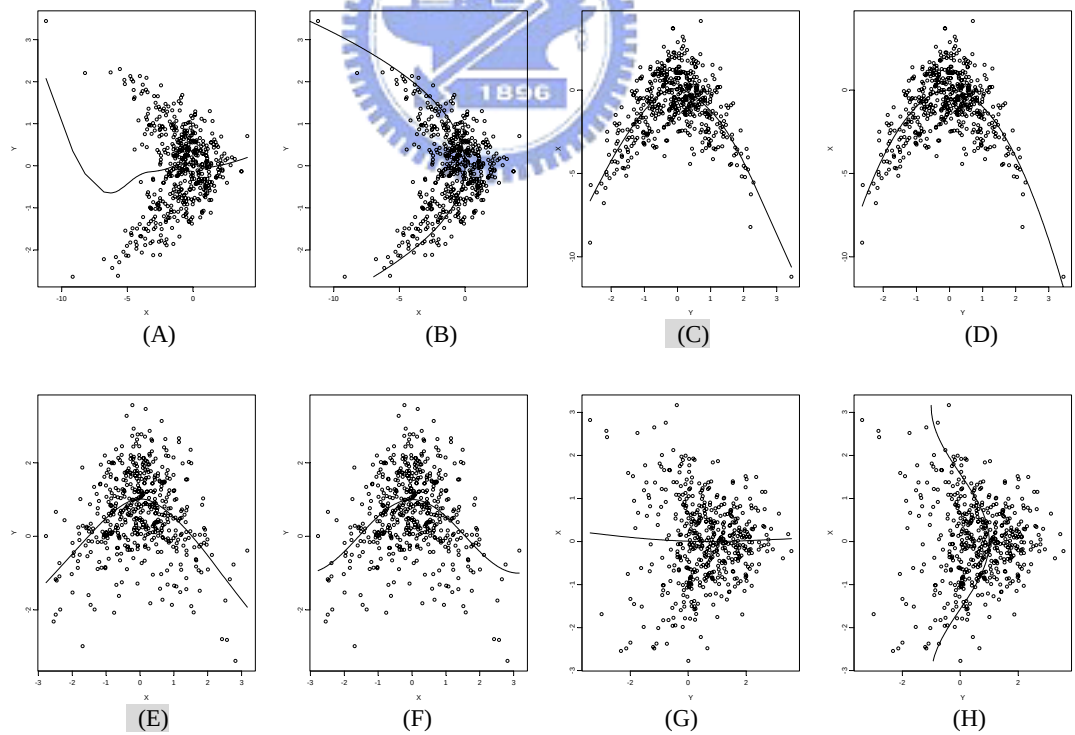


Figure 8: (A) 5→6 fitting function. (B) 5→6 true function. (C) 6→5 fitting function. (D) 6→5 true function. (E) 9→10 fitting function. (F) 9→10 true function. (G) 10→9 fitting function. (H)10→9 true function.

(10) The discovery of the threshold of SCORE from Table 6:

1. The threshold value decreases when the sample size decreases.
2. The decreasing speed of the threshold value is slow, when we use the SCORE method with adjusting lambda.
3. The graph size of SCORE without adjusting lambda is larger than that with lambda adjusted.

4.2 A study on Causal Relationships

1. Data generation:

$X \sim N(0,1),\ X_2 \sim N(2,1),$
$\varepsilon \sim N(0,1),\ \varepsilon_2 \sim$N(0,0.5).
Functions:
$Y = e^X + \varepsilon,\ Y = X_2^{1/3} + \varepsilon_2,\ Y = X_2^{-1/3} + \varepsilon_2,\ Y = X + \varepsilon,\ Y = \sin(X_2) + \varepsilon,\ Y = X^2 + \varepsilon.$
The Sample size (N) = 1000.

2. Analysis Methods: RANK, RANGE, and SCORE.

3. Results:

(1) Our method can detect the monotone functional structures. But if the functions are symmetric, like $f(X) = X^2$, it is not easy to detect. But if we detect that there is a relation between those two variables, there is about 90% correct rate (see Table 9). We suggest not to use 2 as the number of classes of ordinal type data (Table 8 shows the accurate rate may decrease to 0.340/0.328).

(2) We find that using 4 as the number of classes of the ordinal type data gets a better result compared with other number of classes (see Table 7 ~ Table 9).

(3) We can see that if we only consider two variables, SCORE will be better than our methods (Compare Table 8 with Table 10). But if the number of variables is more than two and the sample size is small, our method will be better.

4.3 Real Data Analysis

1. Data source:

S. cerevisiae cell-cycle measurements of Spellman et al. (1998)

http://cell-cycle-www.standford.edu

http://www.molbiolcell.org/

We use the data of the experiment CDC28. The data were collected at 17 different time points.

2. Analysis Methods: The sample size is very small (N=17), we differentiate the relations between genes by RANK, RANGE, and SCORE. We use 0.72 (from Table 6) as the threshold of SCORE.

3. Analysis results: In the following, we show some of the resulting gene network

(1) Case 1:

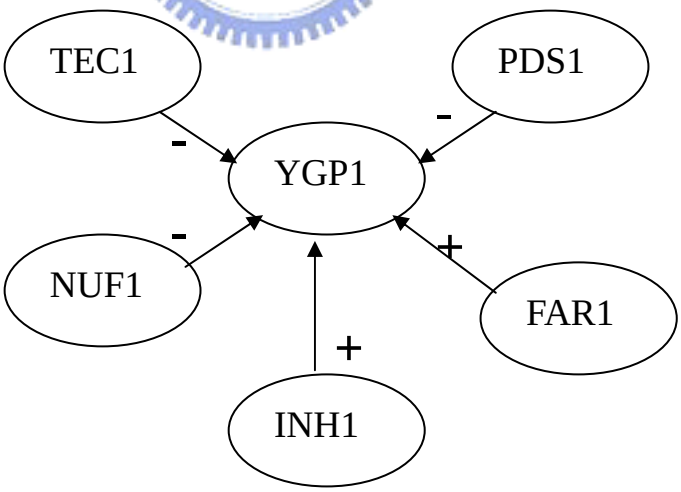


Figure 9: A partial predicted gene regulatory network for the yeast data (CDC28) from Xu et al. (2002). This network is constructed not only by the Smooth Response Surface algorithm but also by the exiting biological knowledge.

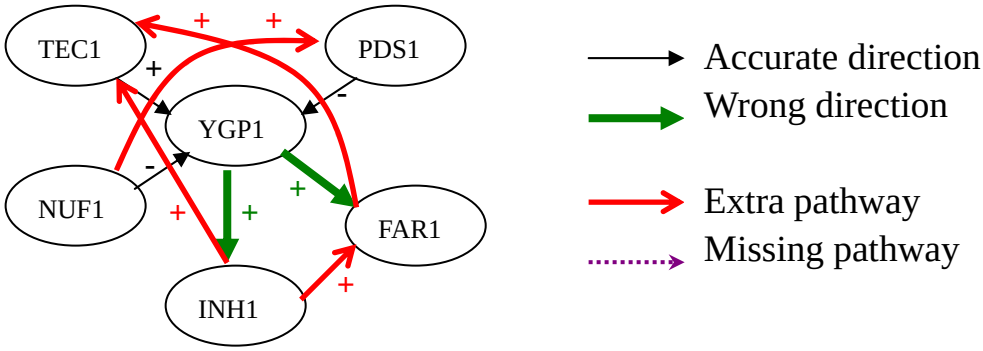


Figure10: The resulting network constructed by SCORE
The threshold of score is 0.72

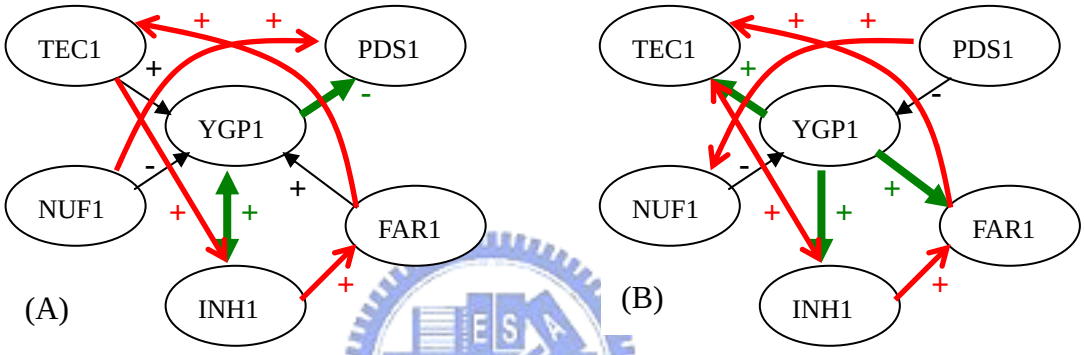


Figure 11: The resulting network constructed by our methods.
(A) RANK with K=4 and without adjusting lambda.
(B) RANGE with K=4 and without adjusting lambda.

The graph size of RANK and RANGE are both 9 and equal to SCORE after being adjusted. Their graph sizes are similar if we use 0.72 as the threshold of SCORE.

We can see that the difference of adjusting lambda or not is small when using RANGE. Comparing with Figure 9, it seems that the number of wrong pathways are less with K = 2.

For RANK with K = 4, without adjusting lambda, it is more similar to Figure 1 than that with lambda adjusted (See Figure 18 (A) and Figure 18 (C)). But using RANK with K = 2 will do better with lambda adjusted (See Figure 18 (E) and Figure 18 (G)). This may be due to that the Yates' continuity correction will be applied when the

expected cell counts of chi-square test are smaller than five. Chi-square test's validity depends heavily on the assumption that the expected cell counts are at least moderately large; a minimum size of five is often quoted as a rule of thumb. In both RANK and RANGE, the number of two-way pathways is large with K = 2 (See Figure 18).

Case 2:

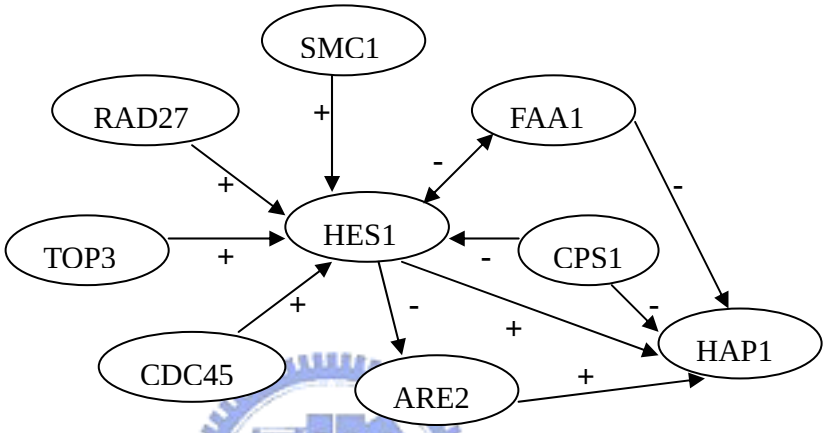


Figure12: A partial predicted gene regulatory network for the yeast data (CDC28) from Xu et al (2002). This netwrok was not only constructed by Smooth Response Surface algorithm but also adjusted by the exiting biological knowledge.

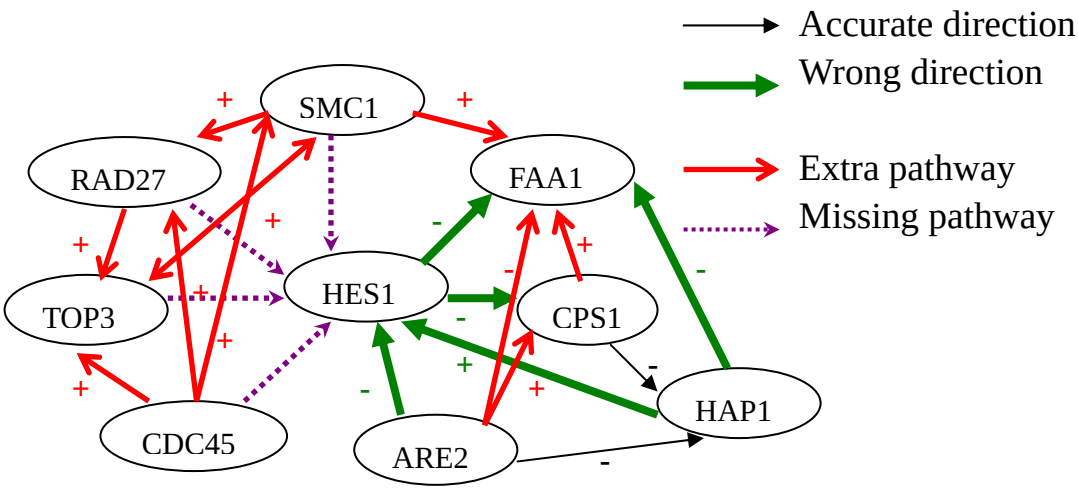


Figure 13: The figure of SCORE when lambda adjusted. The threshold of score is 0.72 (from the Monte Carlo simulation).

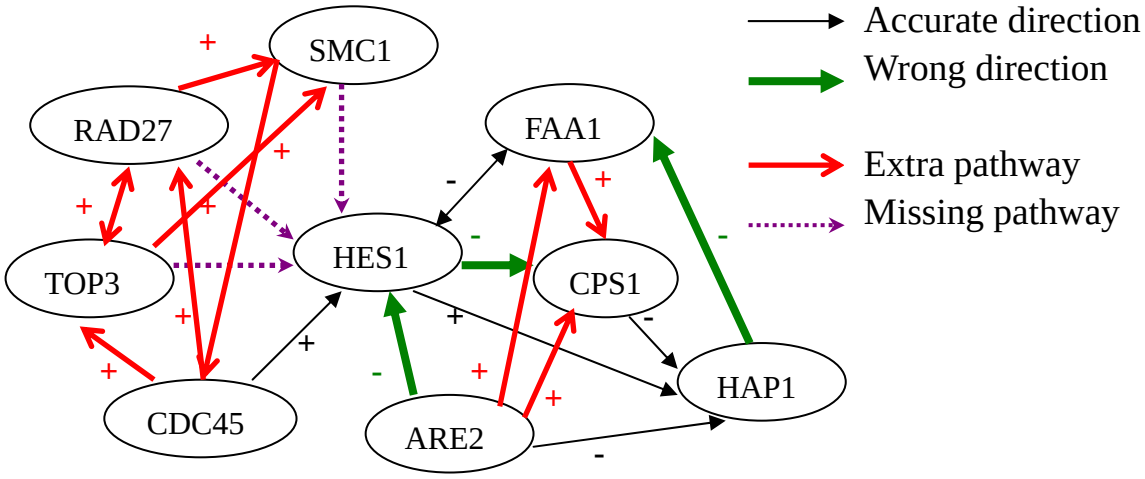


Figure14: RANK with K=4 without adjusting lambda.

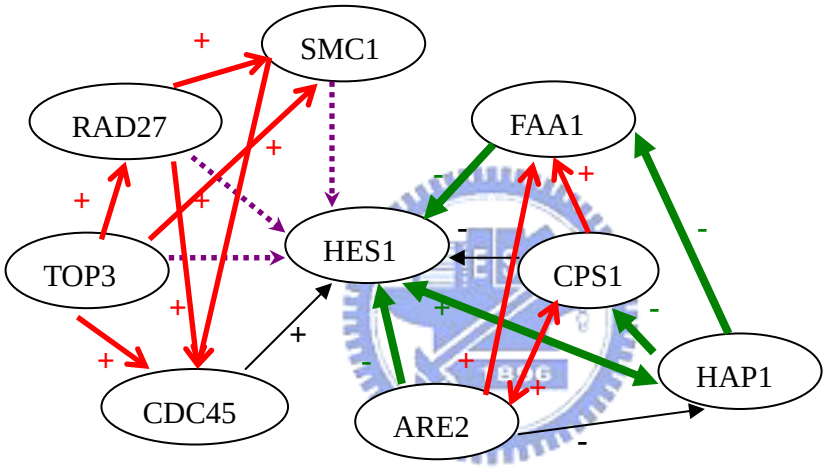


Figure15: RANGE with K=2 and without adjusting lambda.

In this case we can find that the graph size of SCORE is similar with that of RANK and RANGE. But if we use SCORE and do not adjust lambda (Figure 20 (A)), the graph size of SCORE will be larger than that of our methods. We find that using RANK and K=4 will do better when without adjusting the lambda, RANGE and K=4 will do better when lambda adjusted. There is no big difference when K=2 whether adjusting lambda or not. Summarizing the above results, RANK with K=4 is closer to Figure 12.

After checking these two cases, we find that RANK with K=4 and without adjusting the lambda is closer to Figure 9 and Figure12.

5. Discussions

The advantages of our method are we don't need strong statistical assumption before using and compute rapidly. Especially when the sample size is small, the resulting graph size is still similar to the graph size of larger sample size. We give the following conclusions:

1. The method proposed by Cheng (2003) is not always proper for small sample sizes. We find that the threshold of independence need to be adjusted adapt to the sample size when sample size is smaller than 100. If we all use 0.9 for the threshold score and ignore the effect of the sample size, the graph size will grow quickly as the sample size decreases. We suggest that we should find the threshold score by simulation before starting the real data analysis. Table 4 gives the result of the threshold score from our simulation study.
2. If we only consider *two variables*, the method proposed by Cheng will be better than our method. If we need to identify the relevant connections between *sets of genes*, our method will be better. When the threshold point is not adjusted, the graph size of the result using the method proposed by Cheng will be a serious problem.
3. From simulation results, we can see that the resulting graph sizes of our method do not change acutely when the sample sizes are different.
4. There are some particular functional structures that can not be easily detected by our methods. For example, $f(x) = x^2$. The correlation test of the first step cannot detect some symmetrical functions. But we can adjust the result by the second step. If the functional structure between two variables is symmetrical and we cannot detect from the first step, the p-value of the chi-square test for

one direction at the second step will be very close to zero (<0.0001). The detail is showed in the Table 11.

5. Cheng (2003) has mentioned that if the lambda is too small, then the fitting curve will be too rough. If a lambda is smaller than 10^{-6}, it is adjusted to a new lambda such as $\lambda_{new} = \frac{\lambda_1 + \lambda_2}{2}$ or $\lambda_{new} = \sqrt{\lambda_1 \lambda_2}$. Our methods (RANK and RANGE) also use the nonparametric regression to estimate the curve. But we find that the magnitude of lambda has no big influence for our methods. If we ignore the small value of lambda and do not adjust the lambda, we still can discover the relations. Also, without adjusting lambda, we may get a better direction.

We consider the following problems as future works:

1. How to discretize the continuous data and not miss much of the information.
2. Use other correlation tests which are sensitive to the unusual points. We use two nonparametric rank correlation tests that are both not sensitive to the unusual points. That may ignore the effect of influential points.
3. Our method is not sensitive with symmetric and linear functional structure. How to cope with this situation?

References

[1] Akutsu T, Kuhara S, Maruyama O, and Miyano S. (2003). “Identification of gene regulatory networks by strategic gene disruptions and gene overexpressions.” *Theor. Comput. Sci.*, 298**,** 235-251. (Preliminary version has appeared in *Proc. 9th ACM-SIAM Symp.* “Discrete Algorithms.” (1998), 56, 695-702.)

[2] Akutsu T, Miyano S, and Kuhara S. (1999). “Identification of genetic networks from a small number of gene expression patterns under the Boolean network model.” *Pac. Symp. Biocomput.*, 4, 17-28.

[3] Akutsu T, Miyano S, Kuhara S. (2000). “Algorithms for identifying Boolean networks and related biological networks based on matrix multiplication and fingerprint function.” *J Comput Biol.*, 7, 331-343.

[4] Akutsu T, Miyano S, Kuhara S. (2000). “Inferring qualitative relations in genetic networks and metabolic pathways.” *Bioinformatics*, 16(8), 727-734.

[5] Chakravarti, Laha, and Roy. (1967) *Handbook of Methods of Applied Statistics*. Volume I, John Wiley and Sons.

[6] Chen T, He HL, Church GM. (1999). “Modeling gene expression with differential equations.” *Pac Symp Biocomput.*, 29-40.

[7] Cheng C Y. (2003). “Determining the Cause-Effect Relationship between Two Variables by Nonparametric Regression.” Master Thesis. Institute of Statistics, National Chiao-Tung University.

[8] Lu C H. (2003). “Determine the causal relationship between two variables.” Master Thesis. Instituteof Statistics, National Chiao-Tung University.

[9] De Hoon M J, Imoto S, Kobayashi K, Ogasawara N, Miyano S. (2003). “Inferring gene regulatory networks from time-ordered gene expression data of Bacillus subtilis using differential equations.” *Pac Symp Biocomput.*, 17-28.

[10] Friedman N. and Goldszmidt M. (1998). *Learning Bayesian Networks with local Structure*. Jordan, M.I. (ed.), Kluwer Academic Publisher.

[11] Friedman N, Linial M, Nachman I, Pe'er D. (2000). “Using Bayesian networks to analyze expression data.” *J Comput Biol.* 7(3-4), 601-620.

[12] Helsel D R and Hirsch R M. (1995). *Statistical Methods in Water Resources, Studies in Environmental Sciences 49*. Elsevier, Amsterdam.

[13] Hogg R V and Tanis E A. (1997). *Probability and Statistical Inference*. MacMillan

Publishing, Co., New York.

[14] Imoto S, Goto T and Miyano T. (2002). “Estimating of genetic networks and functional structures between genes by using Bayesian networks and nonparametric regression.” *Proc. Pacific Symposium on Biocomputing*, 7, 175-186.

[15] Imoto S, Higuchi T, Goto T, Tashiro K, Kuhara S and Miyano S. (2003). “Combining Microarrays and Biological Knowledge for Estimating Gene Networks via Bayesian Networks.” Technical Report, Human Genome Center, Institute of Medical Science, University of Tokyo, Japan.

[16] Imoto S, SunYong K, Goto T, Aburatani A, Tashiro K, Kuhara S, and Miyano S. “Bayesian network and nonparametric heteroscedastic regression for nonlinear modeling of genetic network.” *Journal of Bioinformatics and Computational Biology*, in press. (Preliminary version has appeared in Proc. IEEE Computer Society Bioinformatics Conference.2002, 219-227.)

[17] Jeffrey S Simonoff. (1996). *Smoothing Methods in Statistics*. Springer series in statstics.

[18] Maki Y, Tominaga D, Okamoto M, Watanabe S, Eguchi Y. (2001) “Development of a system for the inference of large scale genetic networks.” *Pac Symp Biocomput.*, 446-458.

[19] Pearl J. (2000). *Causality: models, reasoning, and inference*. Cambridge University Press, Cambridge, New York.

[20] Shmulevich I, Dougherty ER, Kim S, Zhang W. (2002). “Probabilistic Boolean Networks: a rule-based uncertainty model for gene regulatory networks.” *Bioinformatics*, 18(2), 261-74.

[21] Sida´k Z. (1968). “On multivariate normal probabilities of rectangles: Their dependence on correlations.” *Ann Math Statist.*, 39, 1425–1434.

[22] Sida´k Z. (1971). “On probabilities of rectangles in multivariate normal Student distributions: their dependence on correlations.” *Ann Math Statist.* 41, 169–175

[23] Siegel S and Castellan N J. (1988). *Nonparametric Statistics*.

[24] Siegel S. (1956). *Nonparametric Statistics*. McGraw-Hill. London.

[25] Snedecor, George W and Cochran, William G. (1989). *Statistical methods*. English Edition, Iowa Stat University Press.

[26] SunYong K, Imoto S and Miyano S. (2003). “Dynamic Bayesian Network and

Nonparametric Regression for Nonlinear Modeling of Gene Networks from Time Series Gene Expression Data." *Proc. 1st International Workshop on Computational Methods in System Biology*. Lecture Note in Computer Science, 2602, Springer-Verlag. 104-113.

[27] Tamada Y, Kim S, Bannai H, Imoto S, Tashiro K, Kuhara S, and Miyano S. (2003). "Estimating gene networks from gene expression data by combining Bayesian network model with promoter element detection." *Bioinformatics*, 19 Suppl 2, II227-II236.

[28] Xu H, Wu P, Wu CF, Tidwell C, and Wang Y. (2002). "A smooth response surface algorithm for constructing a gene regulatory network." *Physiol Genomics*, 11(1), 11-20.

Tables

Table 2: The summary of larger sample size. i.e., Figures of right below the described **compared with Figure 6**.

Sample size	*K*	*Rank* *Rate/missing/extra/wrong*	*Range* *Rate/missing/extra/wrong*	*Score*
1000	*8*	*77.8~82.2%/2/6/2* Figure 22(B)	*80~84.4%/2/6/1* Figure 22(C)	*91%/0/2/2* Figure 22(A)
1000	*8*	*80~82.2%/1/7/2* Figure 23(B)	*82.2%~84.4%/1/7/1* Figure 23(C)	*91%/2/1/1* Figure 23(A)
500	*8*	*80%/**0**/7/2* Figure 24(B)	*80%/**0**/7/2* Figure 24(C)	*91%/0/3/2* Figure 24(A)
200	*8*	*80%/**2**/7/2* Figure 25(B)	*80%/**2**/7/2* Figure 25(C)	*88.9%/0/3/2* Figure 25(A)
	4	*84.4%/**2**/6/0* Figure 25(D)	*84.4%/**2**/6/0* Figure 25(E)	
100	*8*	*82.2%/**3**/4/1* Figure 26(B)	*80%/**3**/4/2* Figure 26(C)	*86.7%/1/3/2* *S=0.9* Figure 26(A)
	4	*80%/**3**/4/2* Figure 26(D)	*75.6%/**3**/4/4* Figure 26(E)	
100	*8*	*75.6%/**3**/6/2* Figure 27(C)	*77.8%/**3**/6/1* Figure 27(D)	*65%/1/12/3* *S=0.9* Figure 27(A)
				84.4%/1/4/2 *S2=0.73* Figure 27(B)

K: the number of class of ordinal data.

Rate: the similar rate compared with Figure 1. If both the direction and the connected

way (repressive or activate) are the same with Fugure 1, then we say that is correct.

Missing: the number of missing pathways compare with Figure 1.

Extra: the number of extra pathways compared with Figure 1.

Wrong: the number of wrong pathway directions compared with Figure 1.

S= 0.9, the threshold of SCORE.

S2= the threshold decided by consider Figure 1 (adjusted by us).

Table 3: The summary of larger sample size. The resulting graph **compared with Figure 7.**

Sample size	*K*	*Rank* *Rate/missing/extra/wrong*	*Range* *Rate/missing/extra/wrong*	*Score*
1000	*8*	*71.1%/**7**/4/2*	*73.3%/**7**/4/1*	*68.9%/13/0/1*
1000	*8*	*77.8%/**7**/1/2*	*80%/**7**/1/1*	*73.3%/11/0/1*
500	*8*	*75.6%/**7**/2/2*	*75.6%/**7**/2/2*	*75.6%/9/0/2*
200	*8*	*75.6%/**7**/2/2*	*75.6%/**7**/2/2*	*75.6%/9/0/2*
	4	*80%/**7**/2/0*	*80%/**7**/2/0*	
100	*8*	*68.9%/**12**/1/1*	*66.7%/**12**/1/2*	*68.9%/11/1/2*
	4	*66.7%/**12**/1/2*	*62.2%/**12**/1/4*	*S=0.9*
100	*8*	*64.4%/11/3/2*	*66.7%/11/3/1*	*64.4%/8/5/3* *S=0.9*
				71.1%/10/1/2 *S2=0.73*

K: the number of class of ordinal data.

Rate: the similar rate compared with Figure 7. If both the direction and the connected way (repressive or activate) are the same with Fugure 7, then we say that is correct.

Missing: the number of missing pathways compared with Figure 7.

Extra: the number of extra pathways compared with Figure 7.

Wrong: the number of wrong pathway directions compared with Figure 7.

S= 0.9, the threshold of SCORE.

S2= the threshold of Table 6 (adjusted by us).

Table 4: The summary of smaller sample size. Figures of right below the described **compared with Figure 6**.

Sample size	K	lambda	Rank	Range	Score
100	4	N	80%/**2**/6/1 Figure 28(C)	80%/**2**/6/1 Figure 28(D)	77.8%/0/5/5 Figure 28(A)
		Y	82.2%/**2**/6/0		80%/0/4/5 Figure 28(B)
	2	N	80%/**2**/6/1	75.56%/**2**/6/3	
		Y	Figure 28(E)	Figure 28(F)	
50	4	N	80%/**3**/3/3 Figure 29(C)	77.7%/**3**/3/4 Figure 29(D)	82.2%/0/6/2 Figure 29(A)
		Y	75.6%/**3**/3/5 Figure 29(E)	75.6%/**3**/3/5 Figure 29(F)	86.7%/1/4/1 Figure 29(B)
	2	N	75.6%/**3**/3/5 Figure 30(A)	75.6%/**3**/3/5 Figure 30(B)	

		Y	73.3%/**3**/3/6 Figure 30(C)	75.6%/**3**/3/5 Figure 30(D)	
30	4	N	84.4%/**4**/1/2 Figure 31(C)	84.4%/**4**/1/2 Figure 31(D)	64%/2/10/4 Figure 31(A)
		Y	82.2%/**4**/1/3 Figure 31(E)	77.7%/**4**/1/4 Figure 31(F)	71%/0/10/3 Figure 31(B)
	2	N	77.7%/**4**/1/5 Figure 32(A)	82.2%/**4**/1/3 Figure 32(B)	
		Y	80%/**4**/1/4 Figure 32(C)	80%/**4**/1/4 Figure 32(D)	
17	4	N	77.8%/**5**/3/2 Figure 33(C)	75.6%**5**/3/3 Figure 33(D)(F)	73.3%/2/7/3 Figure 33(A)
		Y	80%/**4**/4/1 Figure 33(E)		62.2%/1/12/4 Figure 33(B)
	2	N	75.6%/**5**/3/3 Figure 34(A)	73.3%/**5**/3/4 Figure 34(B)(D)	
		Y	75.6%/**5**/3/3 Figure 34(C)		

N: we ignore whether the value of lambda is too small or not.

Y: If $\lambda_1 < 10^{-6}$ or $\lambda_2 < 10^{-6}$, *then we adjust the lambda by* $\lambda_{new} = \frac{(\lambda_1 + \lambda_2)}{2}$.

Table 5: The summary of smaller sample size. The resulting graph **compared with Figure 7.**

Sample size	K	lambda	Rank	Range	Score
100	4	N	71.1%/10/2/1	71.1%/10/2/1	68.8%/8/1/5
		Y	74.4%/10/2/0		66.7%/6/4/5
	2	N	71.1%/10/2/1	71.1%/10/2/1	
		Y			
50	4	N	62.2%/13/1/3	60%/13/1/4	68.9%/9/3/2
		Y	57.7%/13/1/5	57.7%/13/1/5	71.1%/10/2/1
	2	N	57.7%/13/1/5	57.7%/13/1/5	
		Y	55.6%/13/1/6	57.7%/13/1/5	
30	4	N	62.2%/15/0/2	62.2%/15/0/2	68.8%/7/3/4
		Y	60%/15/0/3	57.8%/15/0/4	75.6%/6/2/3
	2	N	55.6%/15/0/5	60%/15/0/3	
		Y	57.8%/15/0/4	57.8%/15/0/4	
17	4	N	66.7%/13/0/2	64.4%13/0/3	55.6%/12/5/3
		Y	68.9%/13/0/1		62.2%/7/6/4
	2	N	64.4%13/0/3	62.2%/13/0/4	
		Y	64.4%13/0/3		

N: we ignore whether the value of lambda is too small or not.

Y: If $\lambda_1 < 10^{-6}$ or $\lambda_2 < 10^{-6}$, *then we adjust the lambda by* $\lambda_{new} = \frac{(\lambda_1 + \lambda_2)}{2}$.

Table 6: The threshold of SCORE of our simulation.

Sample size	*lambda*	*threshold*
Larger than 200	*Adjust*	*0.9*
100	*Adjust*	*0.73/0.9*
50	*Adjust*	*0.76*
	Without adjusting	*0.76*
30	*Adjust*	*0.76*
	Without adjusting	*0.65*
17	*Adjust*	*0.72*
	Without adjusting	*0.43*

Table 7: The accurate times of 1000 trials.

(/num: the number of two-way directions)

	Step one	Step two					
		RANK			RANGE		
		K=8	K=4	K=2	K=8	K=4	K=2
$Y = e^X$	1000	772	**815**	617/17	387	**735**	668
$Y = X_2^{1/3}$	1000	763	**774**	671/42	761	**854**	620
$Y = X_2^{-1/3}$	1000	**851**	800	717/65	651	**775**	678/3
$Y = X$	1000	486	478	486/71	409	492	479/2
$Y = \sin(X_2)$	994	861	868	**889**/19	783	**912**	843
$Y = X^2$	609	504	**528**	340/48	573	**592**	328

Table 8: The accurate rates of our methods (include step one and step two).

	Our methods					
	RANK			RANGE		
	K=8	K=4	K=2	K=8	K=4	K=2
$Y = e^X$	0.772	**0.815**	0.617	0.387	**0.735**	0.668
$Y = X_2^{1/3}$	0.763	**0.774**	0.671	0.761	**0.854**	0.620
$Y = X_2^{-1/3}$	**0.851**	0.800	0.717	0.651	**0.775**	0.678
$Y = X$	0.486	0.478	0.486	0.409	0.492	0.479
$Y = \sin(X_2)$	0.861	0.868	**0.889**	0.783	**0.912**	0.843
$Y = X^2$	0.504	**0.528**	***0.340***	0.573	**0.592**	***0.328***

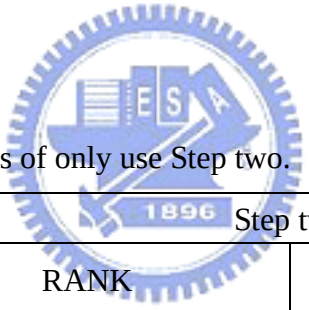


Table 9: The accurate rates of only use Step two.

	Step two					
	RANK			RANGE		
	K=8	K=4	K=2	K=8	K=4	K=2
$Y = e^X$	0.772	**0.815**	0.617	0.387	**0.735**	0.668
$Y = X_2^{1/3}$	0.763	**0.774**	0.671	0.761	**0.854**	0.620
$Y = X_2^{-1/3}$	**0.851**	0.800	0.717	0.651	**0.775**	0.678
$Y = X$	0.486	0.478	0.486	0.409	0.492	0.479
$Y = \sin(X_2)$	0.866	0.873	**0.894**	0.788	**0.918**	0.848
$Y = X^2$	0.828	**0.867**	0.558	0.941	**0.972**	0.539

Table 10: The accurate times of 1000 trials.

	$Y = e^X$	$Y = X_2^{1/3}$	$Y = X_2^{-1/3}$	$Y = X$	$Y = \sin(X_2)$	$Y = X^2$
score	997	988	996	467	989	831

Table 11: N=200, K=8 (“0” is “<0.00001”)

Step one	pair	12	13	14	15	16	17
	p-value of the rank correlation test (Kendall)	0	0	0.06	0	0.11	0
	p-value of the rank correlation test (Spearman)	0	0	0.06	0	0.11	0
	Bonferroni correction (p-value=0.0975)	**0**	**0**	0.1164	**0**	0.2079	**0**
	Kendall’s Tau	0.16	0.12	0.06	0.1	-0.05	-0.52
	Spearman Rho	0.24	0.18	0.08	0.15	-0.07	-0.7
	sign	+	+		+		-
RANK	The p-value of R_1 & X	0	0.54	0.89	0	0.62	0
	The p-value of R_2 & Y	0.78	0.81	0.94	0.77	0.55	0.13
RANGE	The p-value of R_1 & X	0.43	0.6	0.07	0	0.7	0.01
	The p-value of R_2 & Y	0.93	0.53	0.13	0.99	0.76	0.35
SCORE	Lambda1	0.03	5.69	0.25	5.7	5.67	0.02
	S(g1)	0.83	0.97	0.99	0.99	0.99	0.49
	Lambda2	0.02	0.03	0	0	9.31	11.17
	S(g2)	0.72	0.96	0.98	0.82	0.99	0.56

18	19	110	23	24	25	26	27	28	29	210
0	0.51	0.69	0	0	0	0.87	0.14	0.08	0.39	0.26
0	0.49	0.69	0	0	0	0.87	0.15	0.08	0.39	0.27
0	0.7501	0.9039	**0**	**0**	**0**	0.9831	0.269	0.1536	0.6279	0.4598
-0.25	0.02	-0.01	0.26	-0.09	0.13	0	0.04	0.05	0.03	0.03
-0.37	0.03	-0.02	0.39	-0.13	0.19	-0.01	0.06	0.08	0.04	0.05
-			+	-	+					
0.53	0.58	0.56	0.23	0.74	0.54	0.87	0.96	0.14	0.7	0.38
0.08	0.58	0.47	0.78	0.9	0.53	0.92	0.6	0.37	0.36	0.08
0.68	0.78	0.46	0.51	0.13	0.83	0.43	0.95	0.38	0.22	0.29
0.79	0.7	0.75	0.92	0.91	0.05	0.46	0.86	0.36	0.53	0.87
0.02	5.67	0.51	9.07	9.07	9.09	1.29	9.09	0.44	9.07	0.01
0.88	1	1	0.86	0.99	0.97	1	0.99	0.99	1	0.98
0.01	9.65	8.66	0.01	0.07	0.05	9.32	11.18	9.86	9.62	8.64
0.86	1	1	0.83	0.98	0.96	1	0.99	0.99	1	1

34	35	36	37	38	39	310	45	46	47	48
0	0	0.68	0.27	0.13	0.89	0.91	0	0.05	0.03	0
0	0	0.65	0.28	0.12	0.87	0.95	0	0.06	0.03	0
0	**0**	0.888	0.4744	0.2344	0.9857	0.9955	**0**	0.107	0.0591	**0**
-0.15	0.4	0.01	0.03	0.05	0	0	-0.39	-0.06	-0.06	-0.16
-0.23	0.57	0.02	0.05	0.07	-0.01	0	-0.53	-0.08	-0.09	-0.24
-	+						-		-	-
0.04	0.78	0.33	0.36	0.72	0.24	0.77	0.994769	0	0.88	0.95
0.86	0.01	0.35	0.72	0.94	0.42	0.89	0.987262	0	0.29	0.18
0.2	0.79	0.49	0.57	0.27	0.21	0.17	0	0	0.38	0.76
0.17	0.43	0.36	0.63	0.16	0.19	0.13	0.6	0.01	0.35	0.11
0	0.01	9.4	0.53	0	0.49	9.39	0	0	0.04	0
0.77	0.68	1	0.99	0.94	1	1	0.49	0.9	0.99	0.88
0	0	0.17	7.41	0.01	1.84	0.77	0	0	1.59	9.86
0.91	0.62	0.99	0.99	0.98	1	1	0.1	0.27	1	0.97

49	410	56	57	58	59	510	67	68	69	610
0.48	0.86	0.24	0.09	0	0.36	0.53	0.15	0.08	0.07	0.87
0.51	0.86	0.22	0.09	0	0.38	0.54	0.16	0.09	0.06	0.87
0.7452	0.9804	0.4072	0.1719	**0**	0.6032	0.7838	0.286	0.1628	0.1258	0.9831
0.02	-0.01	0.03	0.05	0.09	-0.03	-0.02	0.04	0.05	-0.05	0
0.03	-0.01	0.05	0.08	0.13	-0.04	-0.03	0.06	0.08	-0.08	-0.01
		+		-						
0.73	0.59	0	0.62	0.95	0.15	0.92	0.91	0.57	0.77	0.26
0.14	0.74	0.85	0.64	0.62	0.01	0.87	0.47	0.5	0.91	0.62
0.48	1	0	0.14	0.73	0.65	0.87	0.82	0.9	0.13	0.27
0.28	0.98	0.23	0.46	0.16	0.6	0.98	0.34	0.93	0.6	0.29
0.1	0.1	0	0.09	0.01	0.15	0.58	9.29	0.14	0.03	9.28
0.99	1	0.97	0.99	0.96	0.99	1	0.99	0.99	0.99	1
0.06	0.61	0.01	11.21	9.83	0.01	8.64	1.23	0.16	9.63	0
0.99	1	0.5	0.99	0.98	0.98	1	0.99	0.99	0.99	0.99

78	79	710	89	810	910
0	0.73	0.78	0.87	0.69	0.96
0	0.72	0.75	0.87	0.69	0.96
0	0.9244	0.945	0.9831	0.9039	0.9984
0.42	-0.01	0.01	0	-0.01	0
0.61	-0.02	0.01	-0.01	-0.02	0
0.2	0.87	0.63	0.82	0.93	0.34
0.78	0.89	0.72	0.95	0.8	0
0.12	0.53	0.58	0.75	0.62	0.81
0.78	0.81	0.82	0.55	0.85	0
0.04	0.77	11.21	0.05	9.83	0.02
0.67	1	1	0.99	1	0.77
0.01	9.65	8.64	9.64	8.67	0.6
0.62	1	1	1	1	1
+					+

Figures

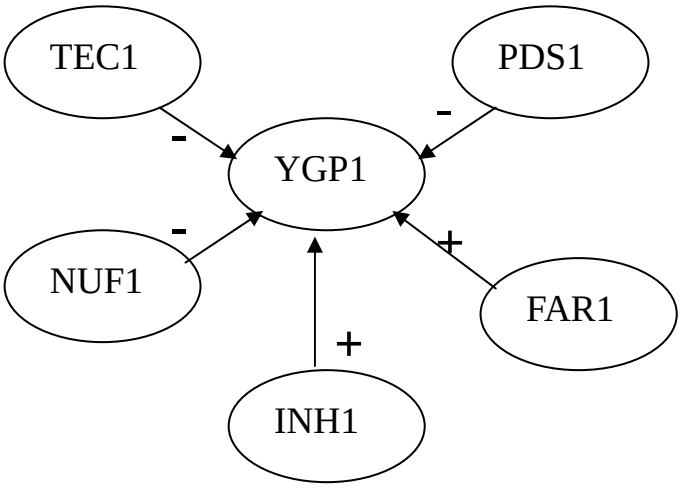


Figure 16: A partial predicted gene regulatory network for the yeast data (CDC28) from Xu et al. (2002). This netwrok is constructed not only by the Smooth Response Surface algorithm also the exiting biological knowledge.

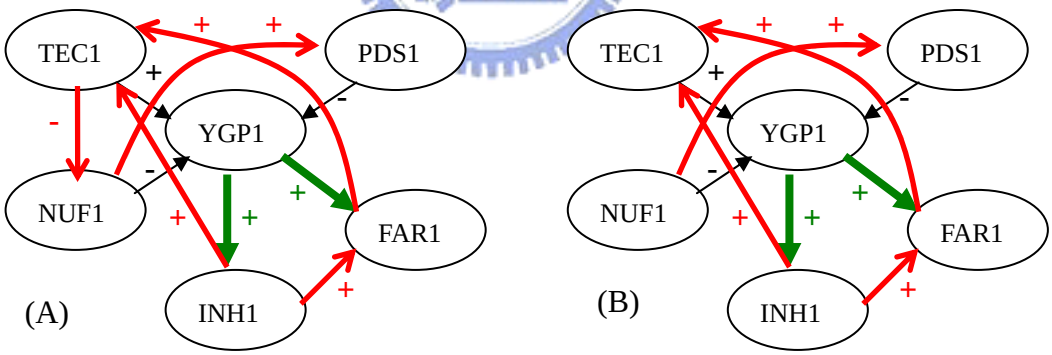


Figure 17: The resulting network constructed by SCORE
(A) Without adjusting lambda. The threshold score is 0.45.
(B) When adjusting lambda. The threshold score is 0.72

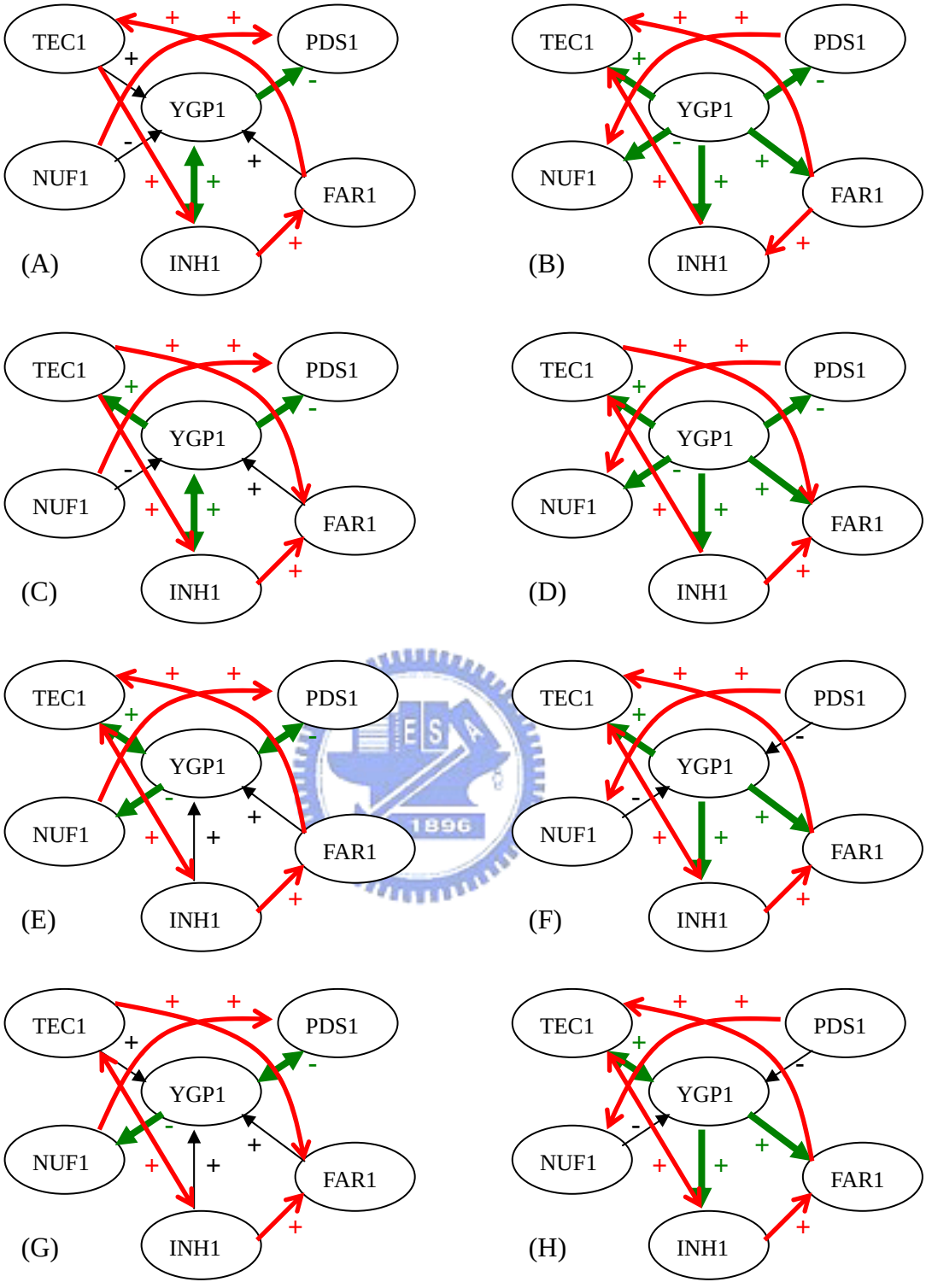


Figure 18: The resulting network constructed by our methods. (A) RANK with K=4 and without adjusting lambda. (B) RANGE with K=4 and without adjusting lambda. (C) RANK with K=4 when adjusting lambda. (D) RANGE with K=4 when adjusting lambda. (E) RANK with K=2 and without adjusting lambda. (F) RANGE with K=2 and without adjusting lambda. (G) RANK with K=2 when adjusting lambda. (H) RANGE with k=2 when adjusting lambda.

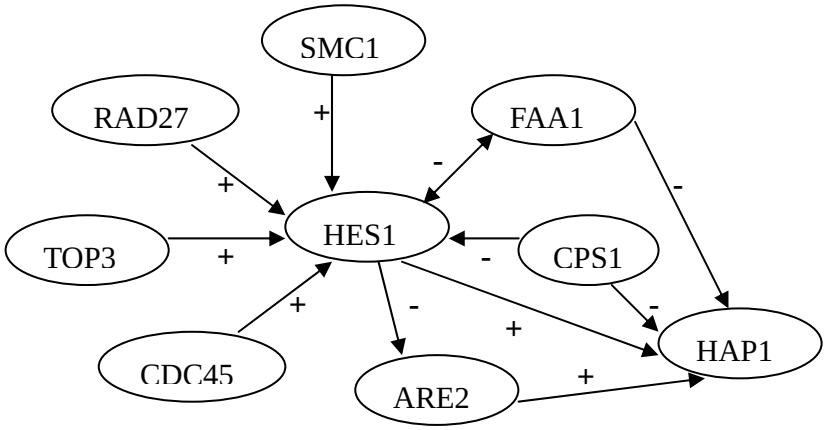


Figure19: A partial predicted gene regulatory network for the yeast data (CDC28) from Xu et al. (2002). This network not only constructed by Smooth Response Surface algorithm but also adjusted by the exiting biological knowledge.

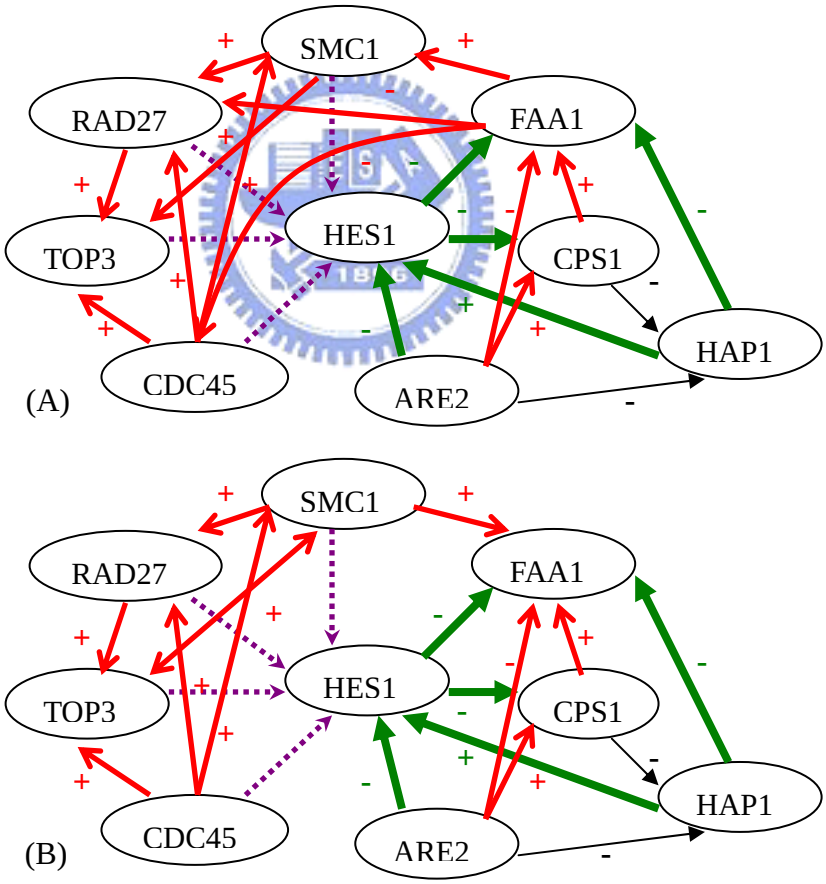


Figure 20: The resulting network constructed by SCORE
(C) Without adjusting lambda. The threshold score is 0.45.
(D) When adjusting lambda. The threshold score is 0.72

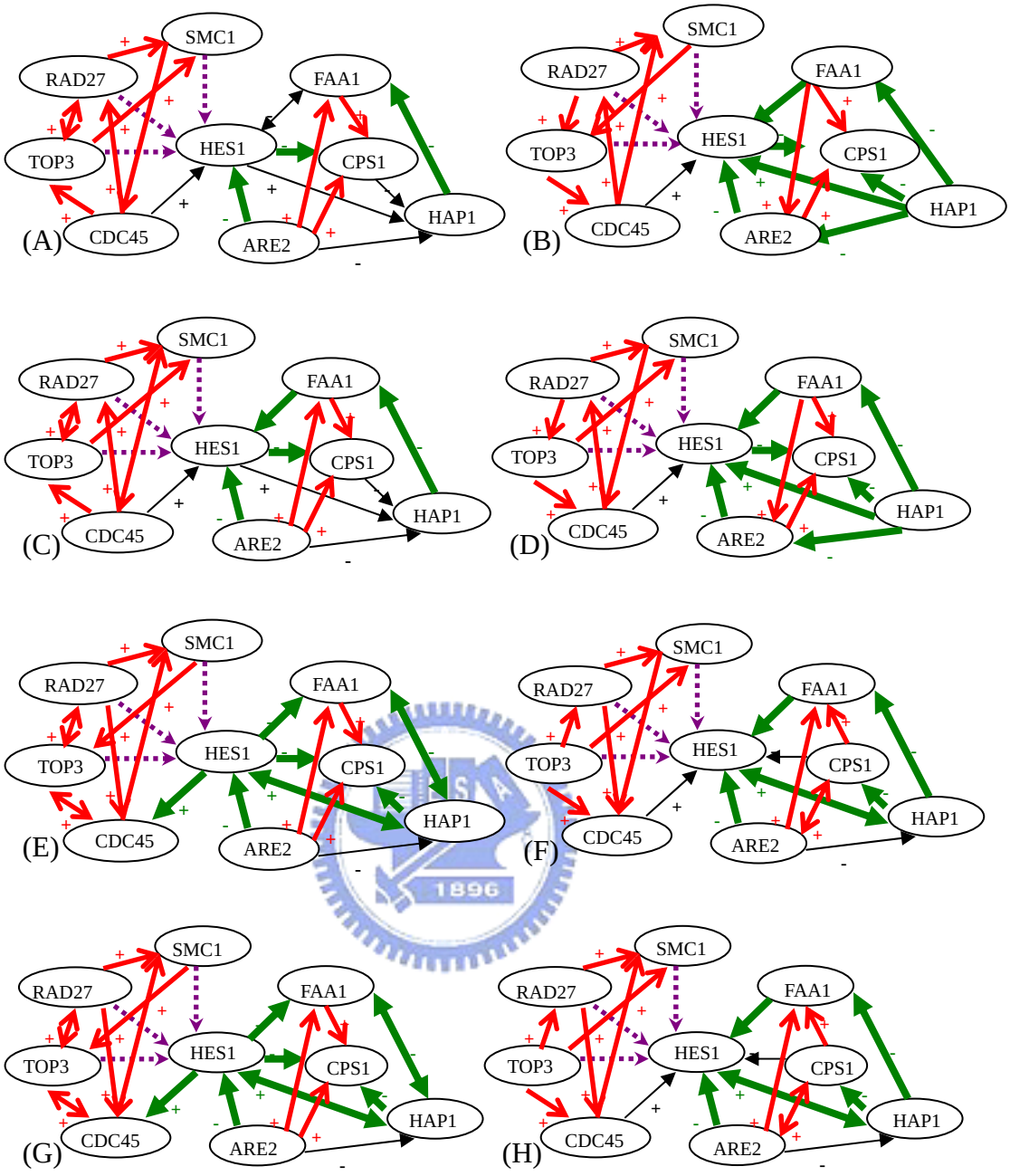


Figure 21: The resulting network constructed by our methods.
(A) RANK with K=4 and without adjusting lambda.
(B) RANGE with K=4 and without adjusting lambda.
(C) RANK with K=4 when adjusting lambda.
(D) RANGE with K=4 when adjusting lambda.
(E) RANK with K=2 without adjusting lambda.
(F) RANGE with K=2 without adjusting lambda.
(G) RANK with K=2 when adjusting lambda.
(H) RANGE with k=2 when adjusting lambda.

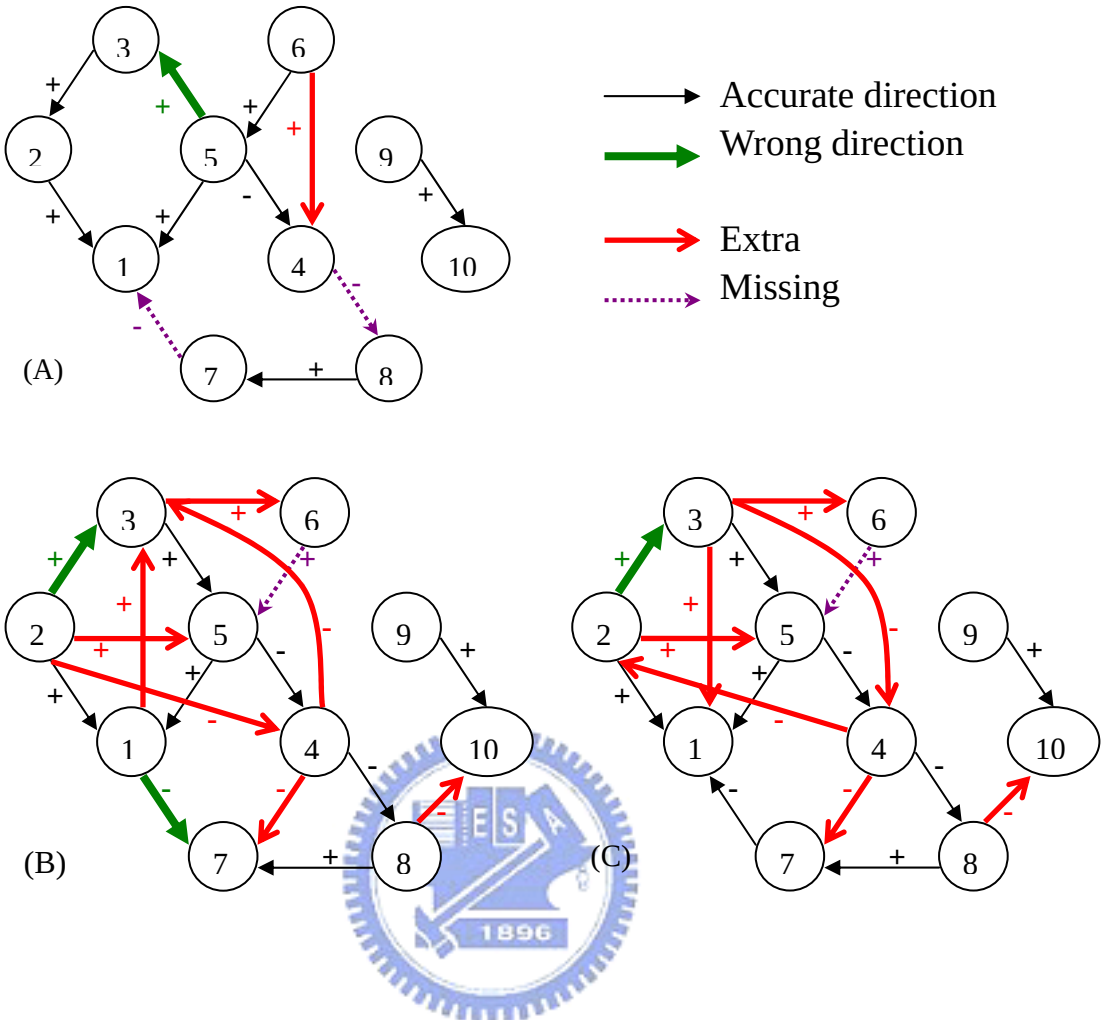


Figure 22: N=1000, K=8 (Table 2)

(A) The result of SCORE**.** (The threshold score = 0.9)

(B) The result of RANK.

(C) The result of RANGE.

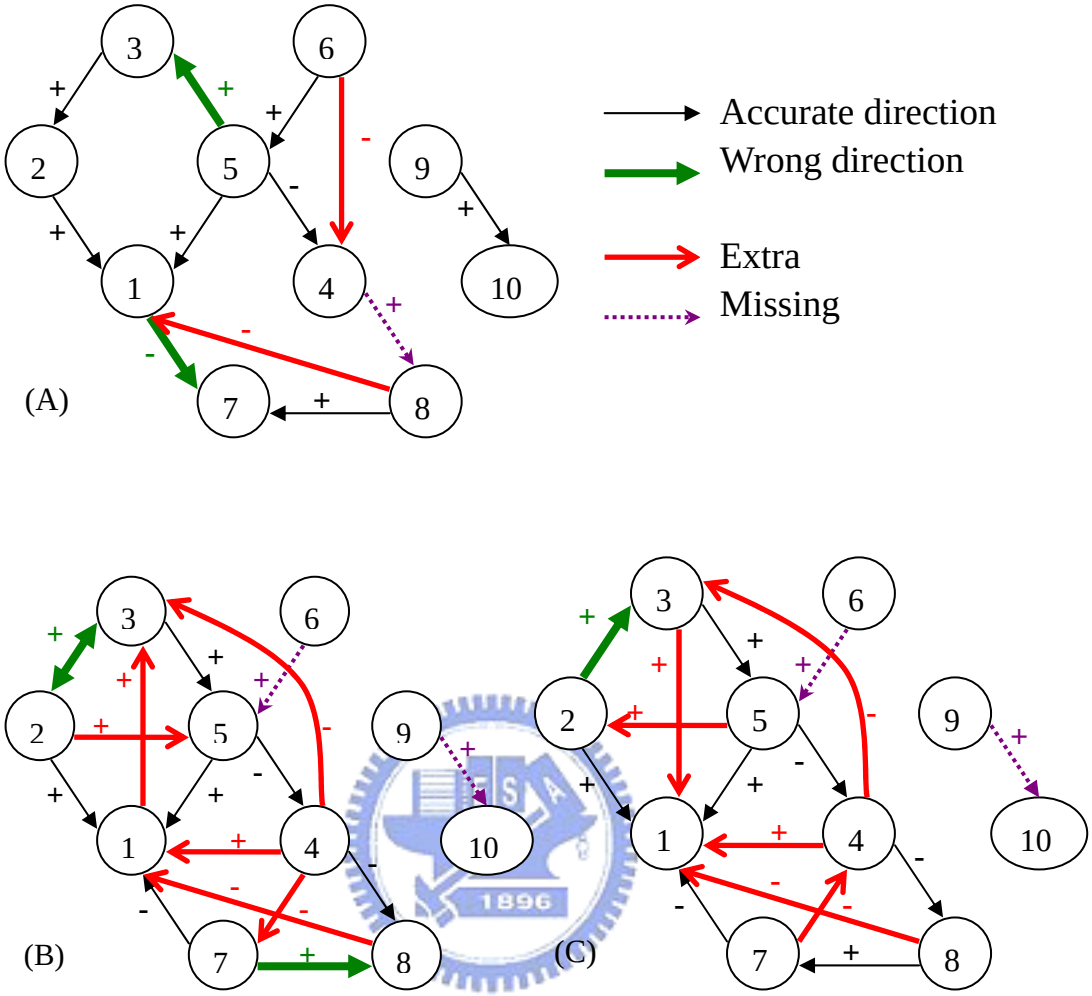


Figure 23: N=1000, K=8 (Table 2)

(A) The result of SCORE**.** (The threshold score = 0.9)

(B) The result of RANK.

(C) The result of RANGE.

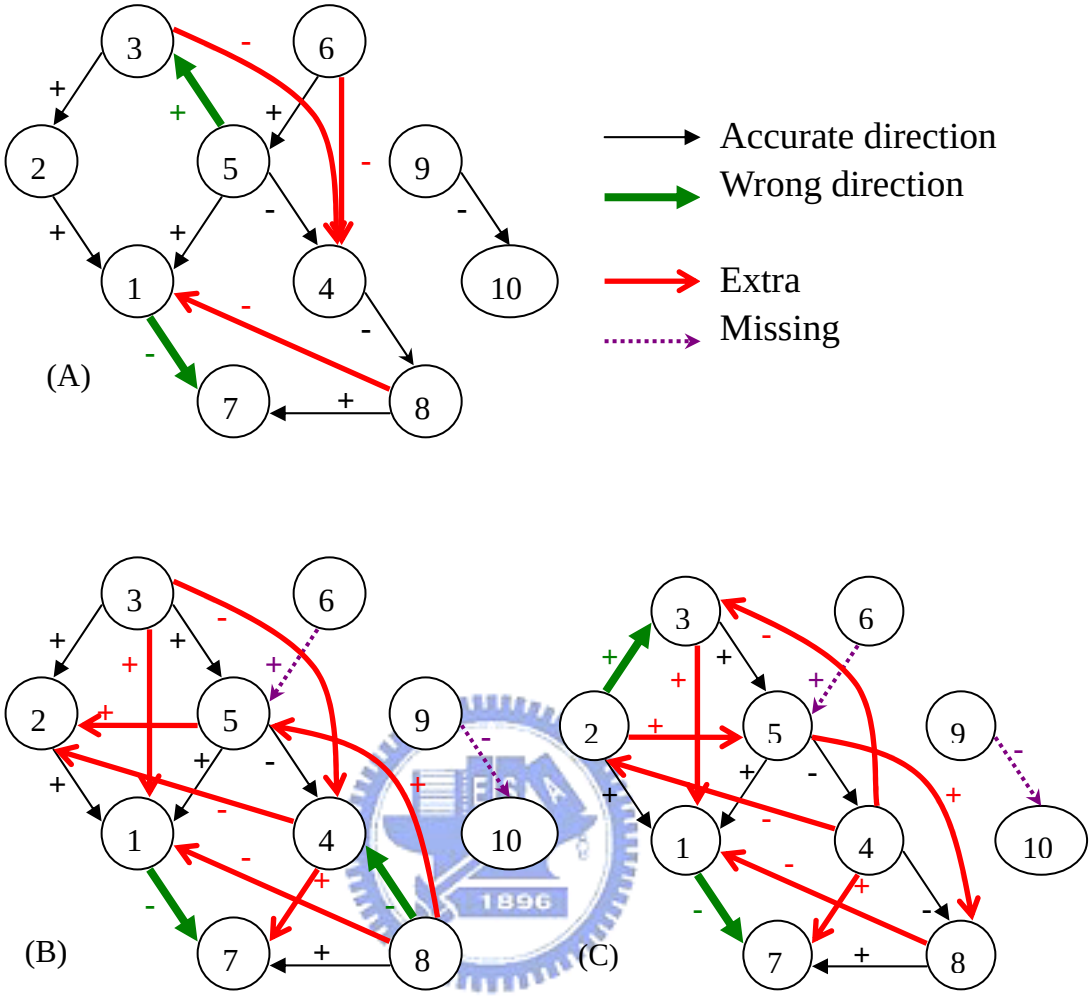


Figure 24: N=500, K=8 (Table 2)

(A) The result of SCORE**.** (The threshold score = 0.9)

(B) The result of RANK.

(C) The result of RANGE.

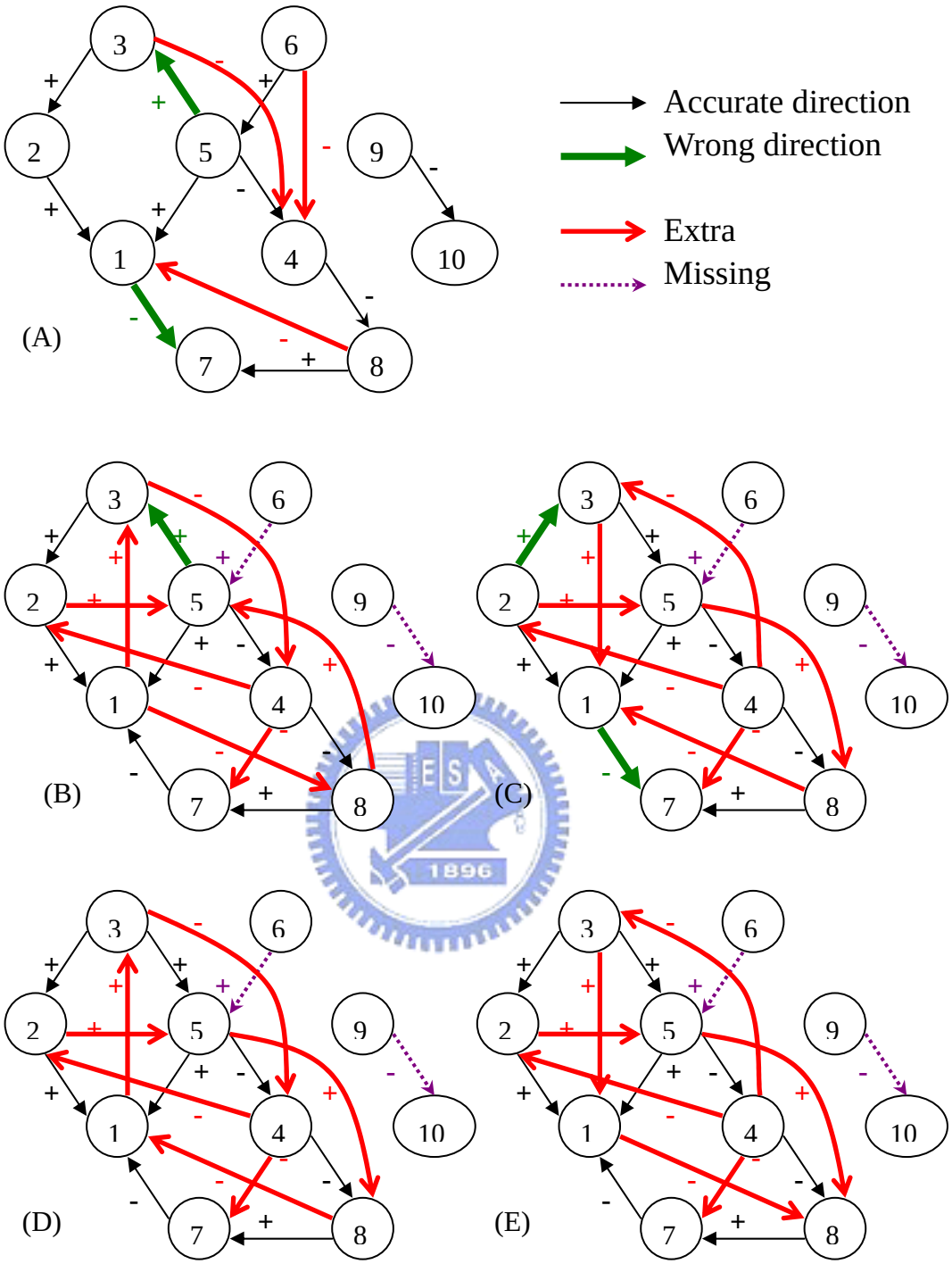


Figure 25: N=200 (Table 2)

(A) The result of SCORE. (The threshold score = 0.9)

(B) The result of RANK with K=8.

(C) The result of RANGE with K=8.

(D) The result of RANK with K=4.

(E) The result of RANGE with K=4.

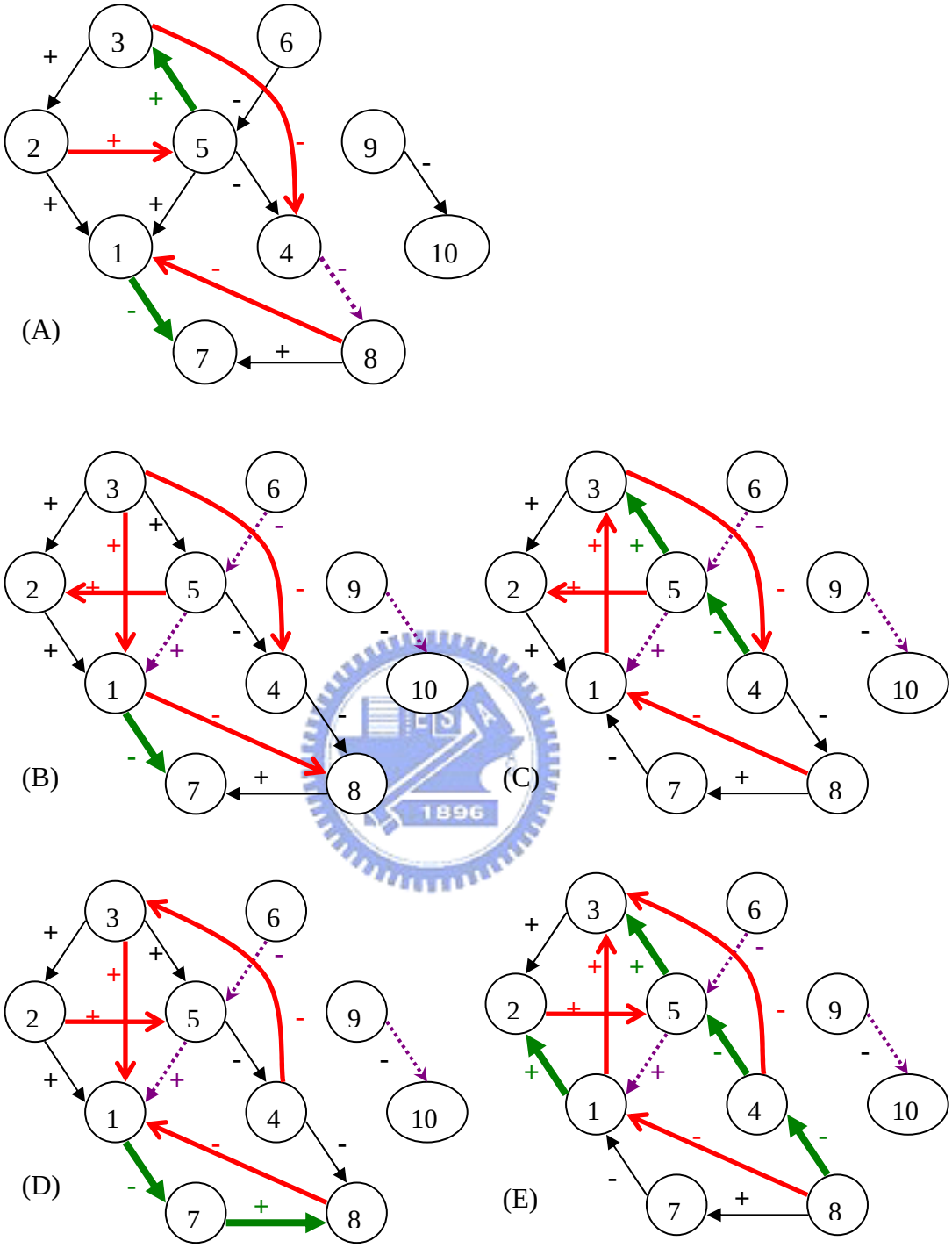


Figure 26: N=100 (Table 2)

(A) The result of SCORE**.** (The threshold score = 0.9)

(B) The result of RANK with K=8.

(C) The result of RANGE with K=8.

(D) The result of RANK with K=4.

(E) The result of RANGE with K=4.

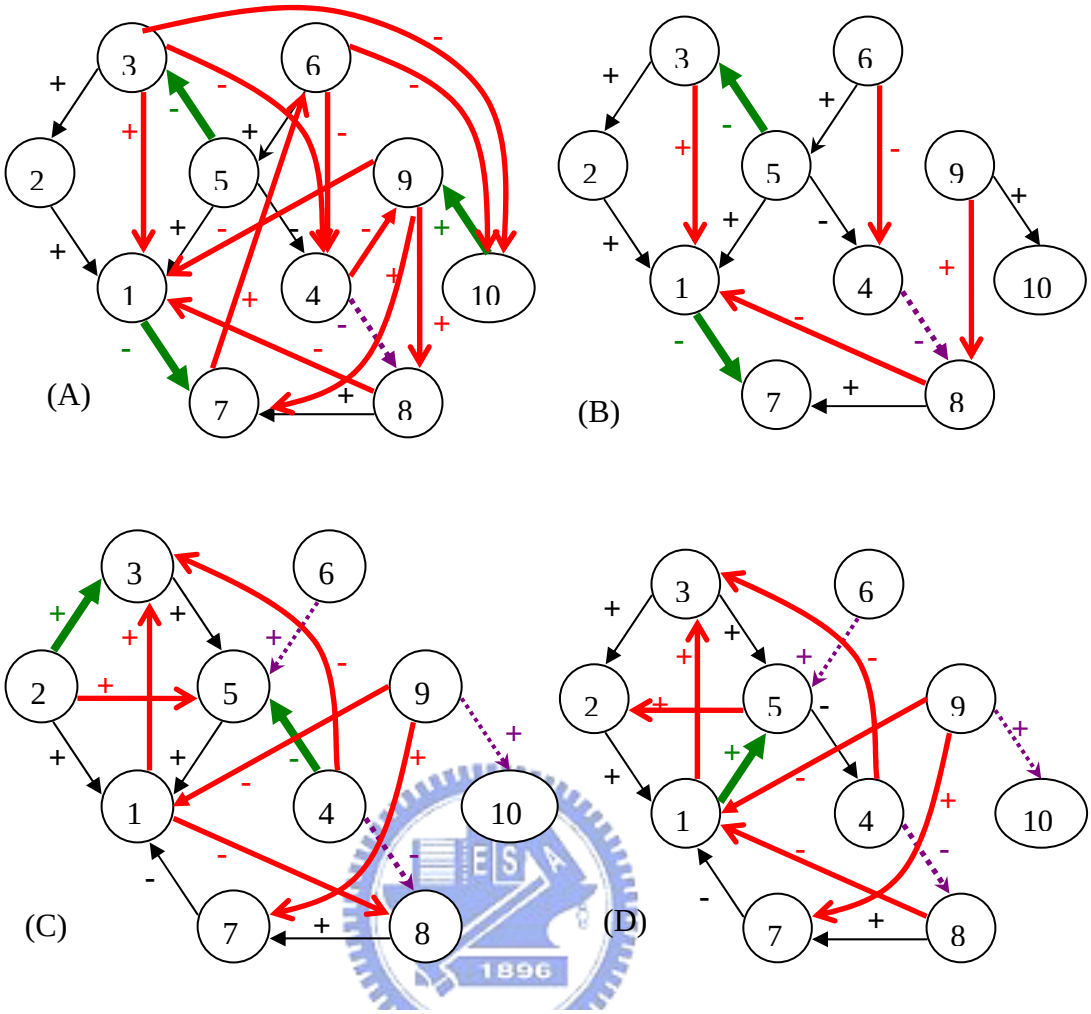


Figure 27: N=100 (Table 2)

(A) The result of SCORE. (The threshold score = 0.9)

(B) The result of SCORE. (The threshold score = 0.73)

(C) The result of RANK with K=8.

(D) The result of RANGE with K=8.

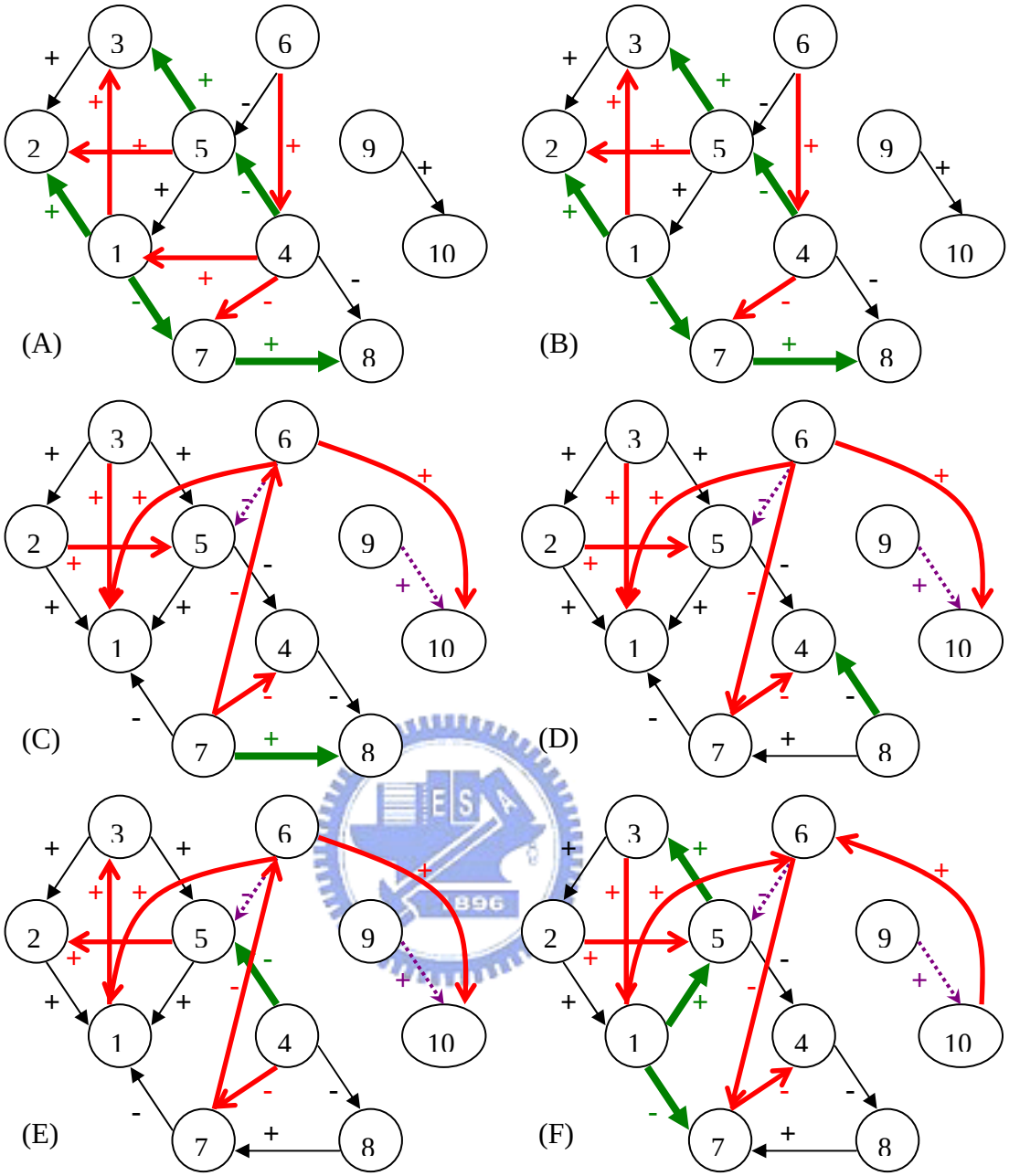


Figure 28: N=100 (Table 4)

(A) The result of SCORE without adjusting lambda**.** (The threshold score = 0.74)

(B) The result of SCORE when adjusting lambda**.** (The threshold score = 0.74)

(C) The result of RANK with K=4.

(D) The result of RANGE with K=4.

(E) The result of RANK with K=2.

(F) The result of RANGE with K=2.

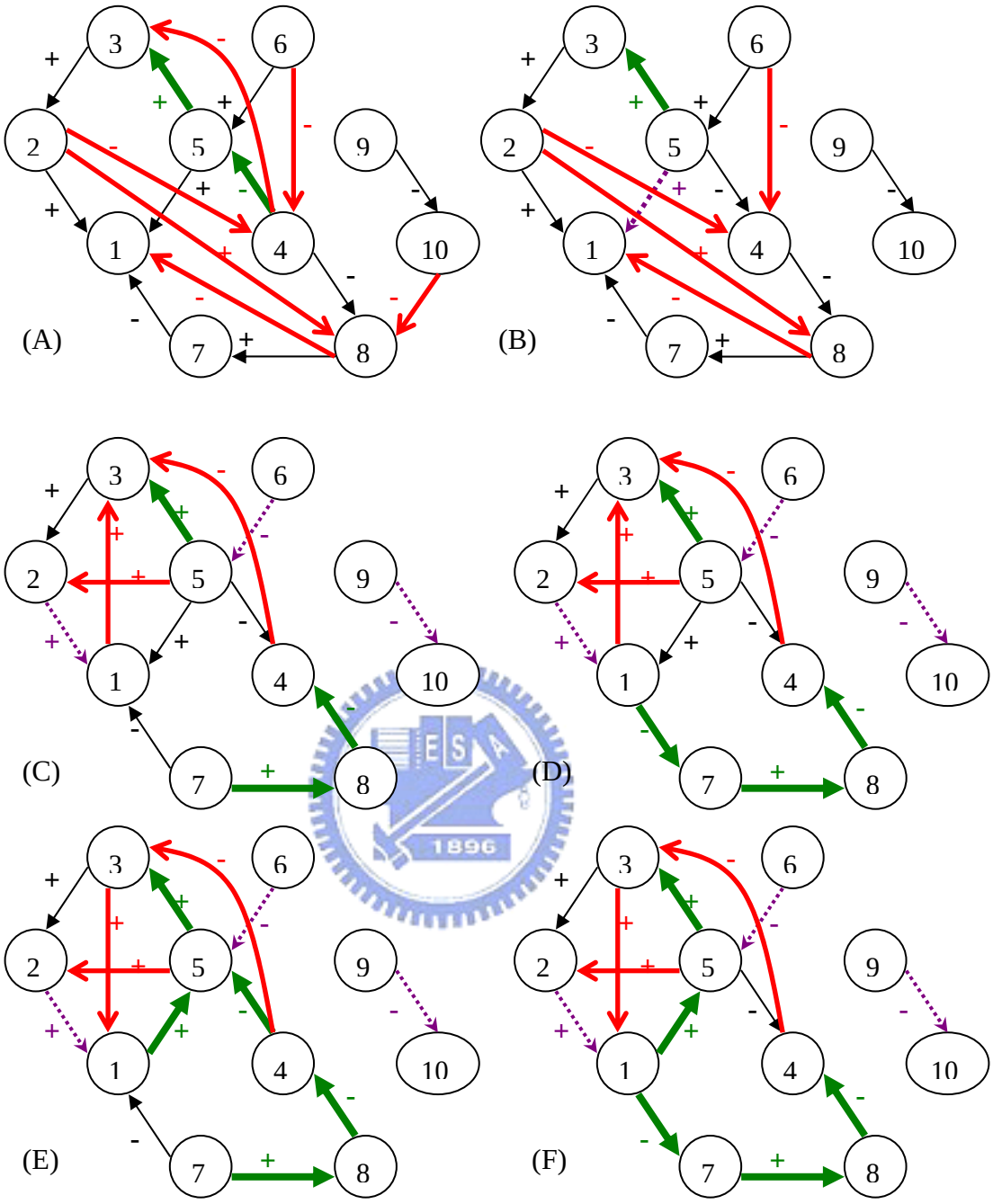


Figure 29: N=50 (Table 4)

(A) The result of SCORE without adjusting lambda**.** (The threshold score = 0.76)

(B) The result of SCORE when adjusting lambda**.** (The threshold score = 0.76)

(C) The result of RANK and K=4 without adjusting lambda.

(D) The result of RANGE and K=4 without adjusting lambda.

(E) The result of RANK and K=4 when adjusting lambda.

(F) The result of RANGE with K=4 when adjusting lambda.

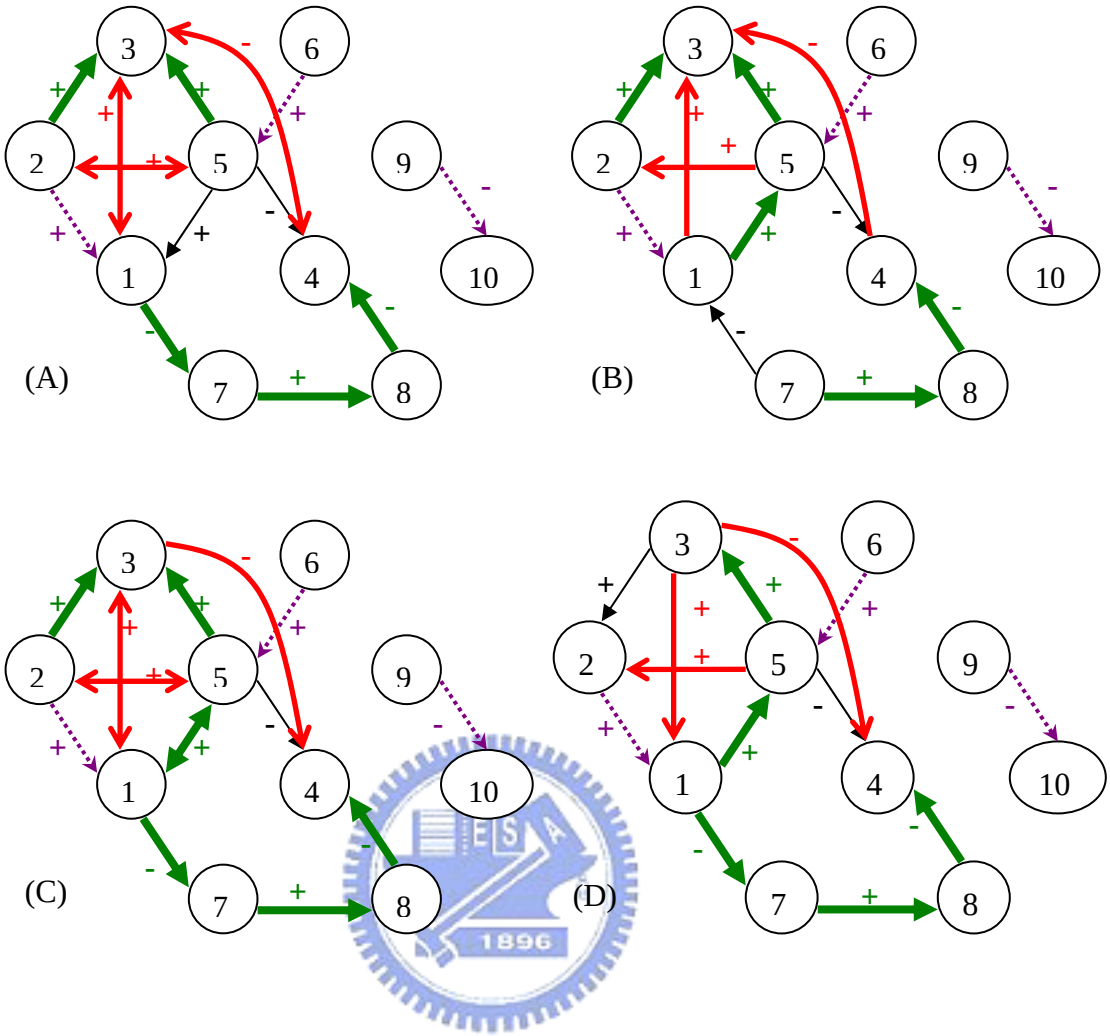


Figure 30: N=50 (Table 4)

(A) The result of RANK and K=2 without adjusting lambda.

(B) The result of RANGE and K=2 without adjusting lambda.

(C) The result of RANK and K=2 when adjusting lambda.

(D) The result of RANGE with K=2 when adjusting lambda.

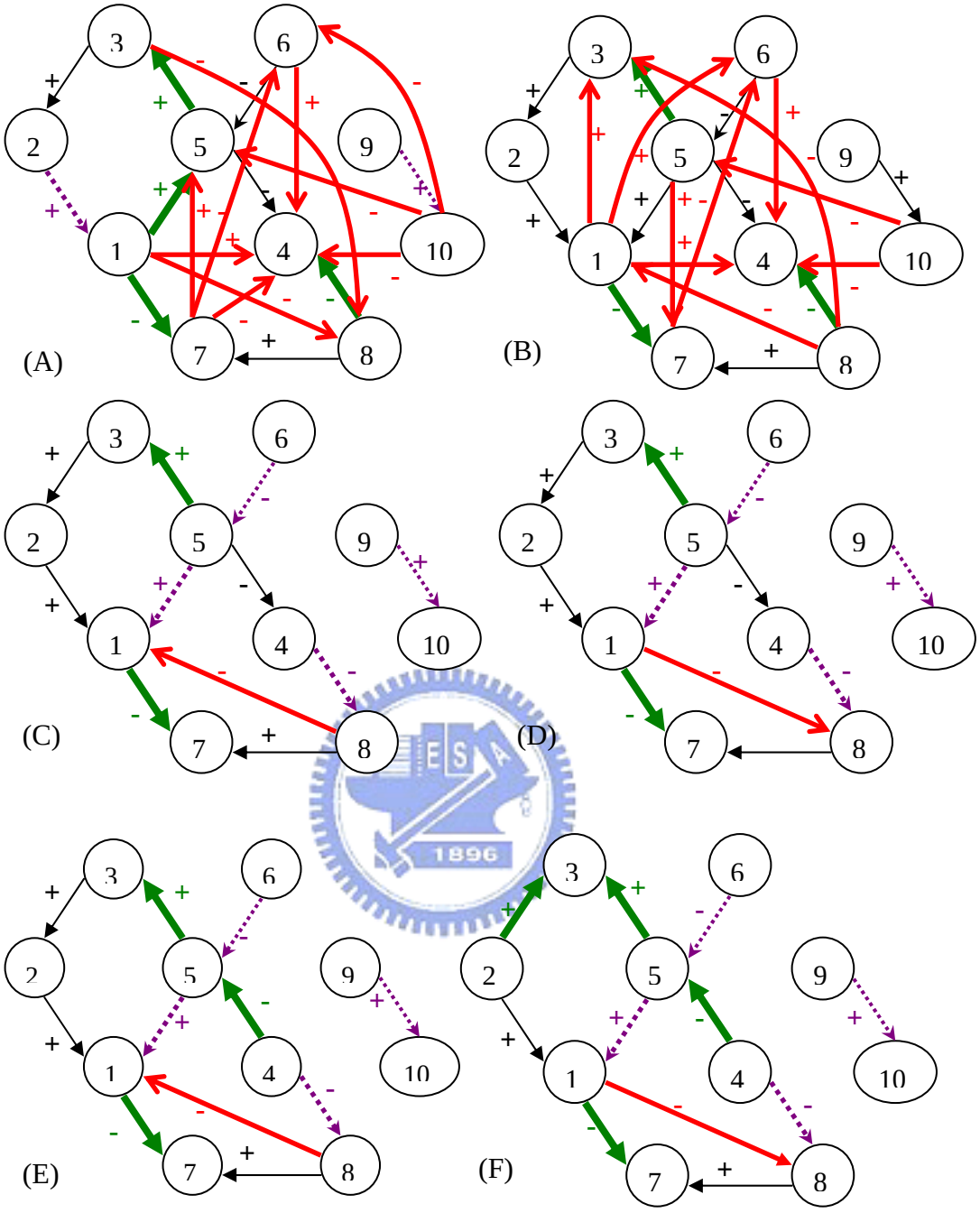


Figure 31: N=30 (Table 4)

(A) The result of SCORE without adjusting lambda**.** (The threshold score = 0.76)

(B) The result of SCORE when adjusting lambda**.** (The threshold score = 0.65)

(C) The result of RANK and K=4 without adjusting lambda.

(D) The result of RANGE and K=4 without adjusting lambda.

(E) The result of RANK and K=4 when adjusting lambda.

(F) The result of RANGE and K=4 when adjusting lambda.

The graph size of SCORE becomes very large compare with our methods.

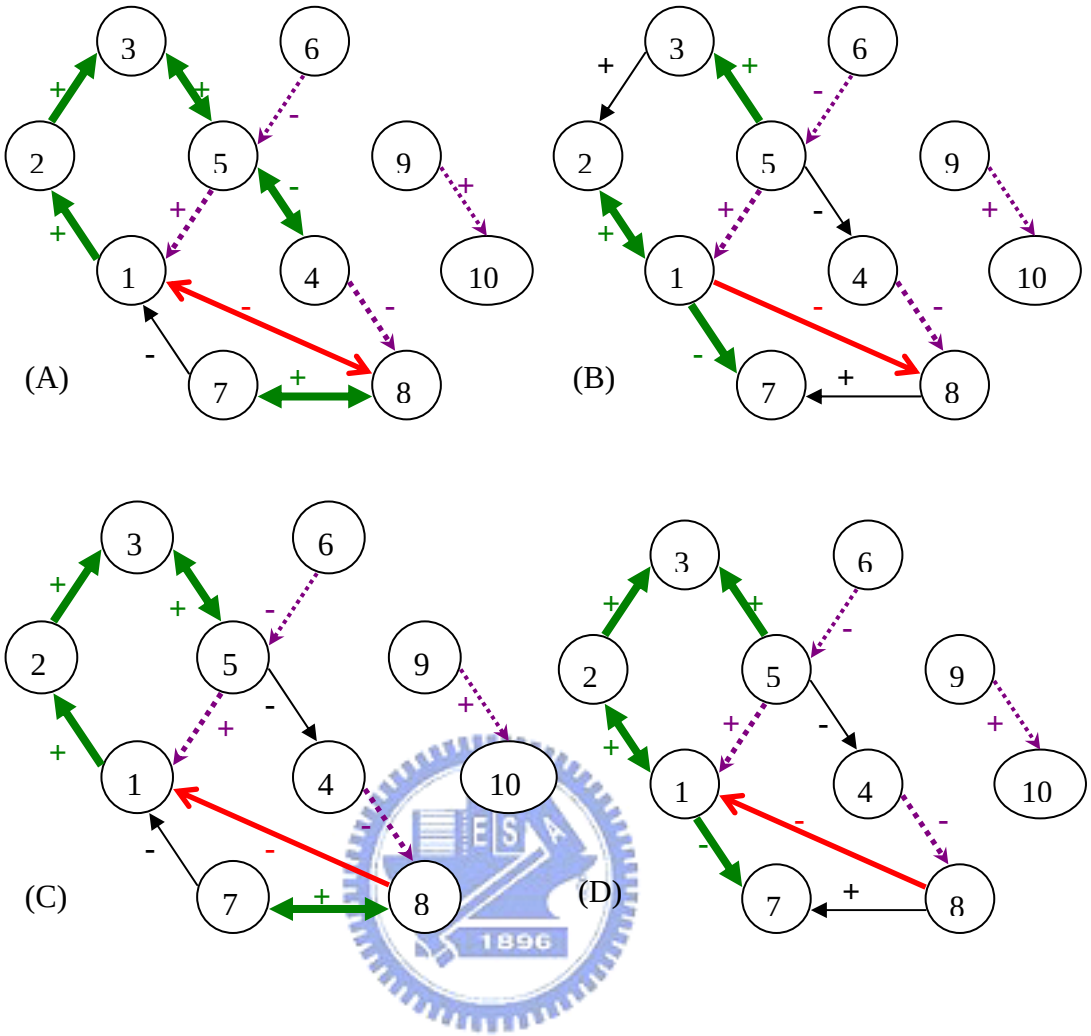


Figure 32: N=30 (Table 4)

(A) The result of RANK and K=2 without adjusting lambda.

(B) The result of RANGE and K=2 without adjusting lambda.

(C) The result of RANK and K=2 when adjusting lambda.

(D) The result of RANGE with K=2 when adjusting lambda.

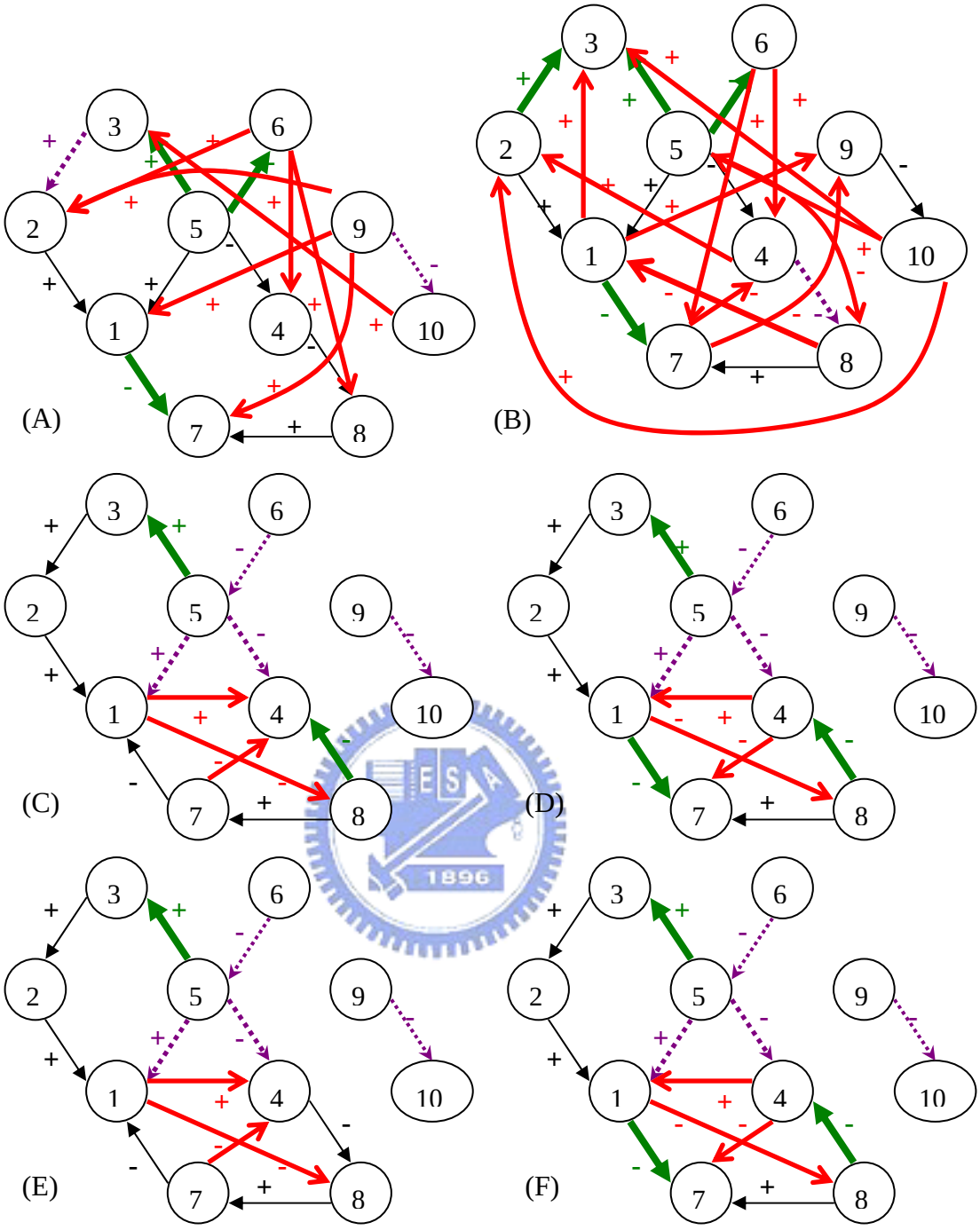


Figure 33: N=17 (Table 4)

(A) The result of SCORE without adjusting lambda**.** (The threshold score = 0.72)

(B) The result of SCORE when adjusting lambda**.** (The threshold score = 0.43)

(C) The result of RANK and K=4 without adjusting lambda.

(D) The result of RANGE and K=4 without adjusting lambda.

(E) The result of RANK and K=4 when adjusting lambda.

(F) The result of RANGE with K=4 when adjusting lambda.

The graph size of SCORE becomes very large compare with our methods.

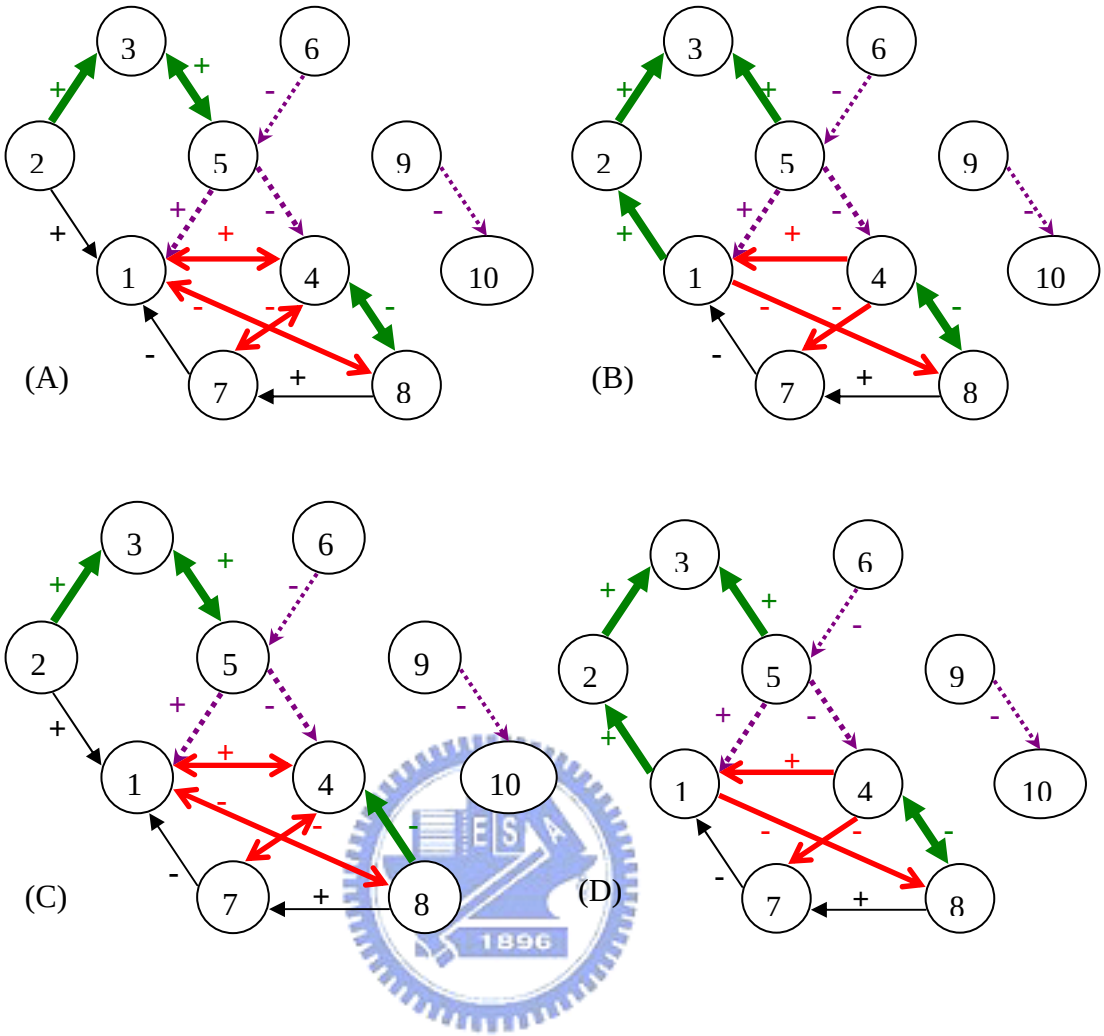


Figure 34: N=17 (Table 4)

(A) The result of RANK and K=2 without adjusting lambda.

(B) The result of RANGE and K=2 without adjusting lambda.

(C) The result of RANK and K=2 when adjusting lambda.

(D) The result of RANGE with K=2 when adjusting lambda.