

國立臺灣大學生物資源暨農學院農藝學研究所  
博士論文



Graduate Institute of Agronomy  
College of Bioresources and Agriculture

National Taiwan University  
Doctoral Dissertation

淨本質相關係數在基因選擇與基因調控  
網路建構之應用

Gene Selection and Regulatory Network Construction with  
Partial Coefficient of Intrinsic Dependence

蕭雅純

Ya-Chun Hsiao

指導教授：劉力瑜博士

Advisor: Li-yu Daisy Liu, Ph.D.

中華民國104年12月

December 2015

國立臺灣大學博士學位論文  
口試委員會審定書

淨本質相關係數在基因選擇與基因調控網路建構之應用  
Gene Selection and Regulatory Network Construction with Partial  
Coefficient of Intrinsic Dependence

本論文係蕭雅純君 (D96621204) 在國立臺灣大學農藝學研究所  
生物統計組完成之博士學位論文，於民國一百零四年十二月十八日承  
下列考試委員審查通過及口試及格，特此證明。

口試委員：

國立臺灣大學農藝學系教授  
廖振鐸 博士

廖振鐸

國立臺灣大學農藝學系副教授  
張孟基 博士

張孟基

私立嶺東科技大學財務金融系助理教授  
歐益昌 博士

歐益昌

國立中興大學農藝學系助理教授  
歐尚靈 博士

歐尚靈

國立臺灣大學農藝學系副教授  
劉力瑜 博士 (指導教授)

劉力瑜



## 謝辭

由衷感謝指導教授劉力瑜老師，從我進入臺大農藝系碩士班一直到博士班，給予我豐富的學習資源與環境，讓我有機會到加拿大參加國際研討會與國外學者交流，並且引領我進入生物資訊的領域，了解實務上生物統計方法的應用。謝謝您這一路上用心的指導與鼓勵，教導我研究知識與態度，在我徬徨失落的時候給予悉心的關懷與包容，讓我有繼續前進的動力。未來我會帶著所學的知識與您的教誨，努力朝下一個階段邁進。

感謝廖振鐸老師在我求學過程中給予我指導與照顧，讓我在碩士班學習到試驗設計方法的研究與應用，並在我博士班研究期間提供很多建議以及給予我很大的鼓勵。感謝張孟基老師、歐益昌老師與歐尚靈老師在我論文口試時細心地審閱論文，給予我很多寶貴的意見。感謝劉仁沛老師在我擔任教學助理期間以及在學期間的關心照顧，讓我有新的視野與教學經驗。同時感謝農藝系生物統計組的所有老師，讓我在求學期間學習到很多不同研究領域的知識，給予我很多溫暖的關心與照顧，讓身為農藝系學生的我真心得覺得很幸福。謝謝詩婷、瑱芳、建郎、柏志、西閔還有所有在研究室一起努力一起歡笑的學弟妹們，謝謝你們豐富了我的研究生生活，給予我充滿樂趣與精采的回憶，我會永遠記得並珍惜與你們一起歡樂的時光。

最後感謝一直默默支持我的家人們，我最敬愛的父親蕭添福先生與母親宋碧銀女士、我最愛的兩位哥哥蕭勝華和蕭勝豪以及我的老公黃彥翔，謝謝你們在我二十幾年的求學歷程上無怨無悔的支持鼓勵我，讓我能全力以赴的在課業上努力，在我開心的時候分享我的喜悅，在我難過時一直陪伴我度過，謝謝你們做我最堅強最溫暖的後盾。

蕭雅純 謹誌

中華民國一百零四年十二月




## 中文摘要

在隨機變數沒有分佈或函數的假設前提之下，本質相關係數依然能夠決定變數間的關係。當計算越多個預測變數與一個目標變數之間的本質相關係數，其數值會越大。這意味著如果存在與目標變數最相關的預測變數且本質相關係數是顯著的，即使再加入其他與目標變數相關性弱的預測變數，其本質相關係數仍然會是顯著的。

在這篇研究當中，我們提出了淨本質相關係數這個方法一步一步地選擇與目標變數相關的預測變數。而且，我們將淨本質相關係數這個方法應用在逐步變數選擇與建構基因調控網路。關於逐步變數選擇的應用，結合本質相關係數與淨本質相關係數這兩個方法可以消除其他相關變數的干擾。從模擬的結果當中，可以觀察到我們所提出的方法比使用結合了皮爾森相關係數與淨相關係數的方法更能具體地發現變數間曲線與直線的關係。根據結合本質相關係數與淨本質相關係數這兩個方法的數值結果，上述的特性提供了指示不同曲線關係程度的機會。在使用公開取得的資料庫之試驗結果中，結合本質相關係數與淨本質相關係數這兩個方法的逐步變數選擇程序能夠成功地鑑別出與三個低溫誘導因子相關的低溫反應基因，並且能有效地辨別樣本相關基因之間的相互作用。因此，我們所提出的策略可能有益於整合分析，並從雜訊中鑑別出相關性的形式。

另一方面，關於建構基因調控網路的策略，使用結合本質相關係數與淨本質相關係數這兩個方法可以在消除被選擇之相關節點的干擾之下，逐步選擇出目標節點與相對應的起始節點。由於本質相關係數與淨本質相關係數的數值具有不對稱性，例如： $CID(Y|X)$  不一定等同於  $CID(X|Y)$  以及  $pCID(Y|X_2; X_1)$  不一定等同於  $pCID(X_2|Y; X_1)$ 。所以我們利用此特性去區別出兩個節點之間的方向性。這個研究進行了虛擬的基因網路，以評估在重複100次不同樣本大小的網路之下使用結合本質相關係數與淨本質相關係數這兩個方法的啟發式演算法之表現。我們可以觀



察到當樣本數增加時，重建的基因網路其正確性也會增加。另外將我們提出的策略應用在兩種不同的微陣列資料庫。其中一個是應用在阿拉伯芥中已知的低溫訊息傳遞路徑，此路徑是經由低溫誘導因子去誘發低溫相關基因(COR)，我們提出的策略能夠成功地找出低溫誘導因子與低溫相關基因之間的連結。另一個資料庫是關於稻米中的鹼性-螺旋-環-螺旋家族，在生物學上還未發現它們的基因網路。因此，運用我們提出的策略建構出一個基因調控網路，可以給生物學家一些參考資訊。

綜合上述，結合本質相關係數與淨本質相關係數這兩個方法能夠有效地鑑別出擁有不同型態關係的相關變數。除此之外，具有不對稱性的本質相關係數與淨本質相關係數可以從統計學的觀點辨別變數間的方向性。因此，根據本質相關係數與淨本質相關係數這兩個方法所得到的變數選擇與建構基因調控網路結果，可以讓生物學家在實驗進行之前當作參考的依據。

關鍵字：本質相關係數、淨本質相關係數、逐步變數選擇、基因調控網路。

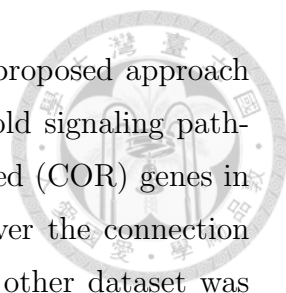


# Abstract

The coefficient of intrinsic dependence (CID) is capable of determining associations among variables without making distributional or functional assumptions regarding to random variables. The CID value of the target variable would increase when more predictor variables include. This implies that a CID value of the target variable given multiple predictors is significant as the most relevant predictor is included even though the other predictors have weak association with the target variable.

In this study, we developed the partial coefficient of intrinsic dependence (pCID) to facilitate the step-by-step selection of variables that are relevant to a target variable. Furthermore, we applied pCID method to stepwise variable selection and the construction of gene regulatory network. In stepwise variable selection, the strategy of selecting relevant variables using the CID along with the pCID can eliminate interference from other relevant variables. From simulation results, we observed that the proposed method is more sensitive to curvilinearity and more specific to linearity than the combination of Pearson's correlation coefficient and the partial correlation coefficient (PCC/pPCC). This property may provide the opportunity to index different levels of curvilinearity according to CID/pCID outcomes. While being exercised on publicly available microarray data, the CID/pCID procedure successfully identified cold-responsive genes related to three C-repeat binding factors, and was especially effective at identifying some sample-specific gene-gene interactions. Therefore, the proposed strategy may be beneficial in meta analysis to distinguish general forms of relationships from the noise.

On the other hand, the strategy of constructing the gene regulatory network using the CID/pCID can stepwise choose the target node and decide the corresponding source node while eliminating the influence of the other relevant nodes. Because of the asymmetric CID/pCID values, we used this property to discriminate the direction of two nodes. Pseudo network was conducted to evaluate the performance of the heuristic approach by CID/pCID from one hundred replications with different sample sizes. As the sample size increased, the accuracy of the re-



constructive pseudo network would increase. Furthermore, the proposed approach was applied to two microarray datasets. One was the known cold signaling pathway, C-repeat binding factors would induce a set of cold-regulated (COR) genes in *Arabidopsis*. The CID/pCID approach could successfully discover the connection between C-repeat binding factor and cold-regulated gene. The other dataset was about the basic helix-loop-helix gene family in rice, which network was undiscovered in biology. We constructed the network based on the CID/pCID outcomes to provide the suggestion for biologists.

In summary, the CID/pCID method could efficiently identify the relevant variables which had various types of the association. Besides, the asymmetric CID/pCID values were used to distinguish the direction of two variables from the statistical viewpoints. Therefore, the statistical outcomes of the variable selection and gene regulated network construction based on the CID/pCID method could provide references for biologists before making an experiment on plants.

Key words: Coefficient of intrinsic dependence, Partial coefficient of intrinsic dependence, Stepwise variable selection, Gene regulatory network.



# Contents

<b>1</b>	<b>Introduction</b>	<b>1</b>
<b>2</b>	<b>Partial Coefficient of intrinsic dependence (pCID)</b>	<b>5</b>
2.1	Coefficient of intrinsic dependence (CID) . . . . .	5
2.2	Partial coefficient of intrinsic dependence (pCID) . . . . .	6
2.3	Estimation of CID and pCID . . . . .	8
2.4	Hypothesis test of Independence for CID and pCID . . . . .	11
2.5	The partial Pearson correlation coefficient (pPCC) . . . . .	11
<b>3</b>	<b>Application to stepwise variable selection</b>	<b>13</b>
3.1	The procedure for selecting variables . . . . .	13
3.2	Simulation study . . . . .	15
3.2.1	The results of CID and pCID . . . . .	16
3.2.2	The results of PCC and pPCC . . . . .	21
3.3	<i>Arabidopsis</i> microarray data analysis . . . . .	23
3.4	Discussion . . . . .	31
<b>4</b>	<b>Application to gene regulatory network</b>	<b>33</b>
4.1	Construction of gene regulatory network by CID/pCID . . . . .	34
4.2	Simulation study . . . . .	38
4.3	<i>Arabidopsis</i> microarray data analysis . . . . .	47
4.4	Rice microarray data analysis . . . . .	52
4.5	Discussion . . . . .	55
<b>5</b>	<b>Conclusions</b>	<b>60</b>



References

A The inference of pseudo network

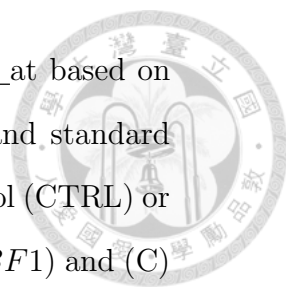
B Supplement table





# List of Figures

3.1	Flow chart of stepwise variable selection based on the CID and pCID.	14
3.2	Boxplots of $pCID(Y X_i; X_4)$ values, $i = 1, 2, 3, 5, 6$ , based 100 simulated samples of size 25, 50, or 100 from the model $Y = 10 \sin(\pi X_1 X_2) + 30(X_3 - 0.5)^2 + 10X_4 + 5X_5 + \varepsilon$ , where $X_i$ 's were distributed as $U(0, 1)$ and $\varepsilon$ was distributed as $N(0, 1)$ . The horizontal line indicates the zero value.	18
3.3	Venn diagrams of the 2,385 cold-responsive genes associated with three CBF transcription factors according to (A) the CID/pCID method, (B) the PCC/pPCC method, and (C) the CID/pCID method and/or the PCC/pPCC method. (D) Venn diagrams of the significantly enriched gene ontology accessions according to the CID/pCID method and/or the PCC/pPCC method.	25
3.4	Expression profiles and CID/pCID inferences of 264052_at and 253534_at based on expression levels of <i>CBF1</i> . (A) Scatter plots of log2 expression levels. (B) Averages and standard deviations of log2 expression levels over time under control (CTRL) or cold treatments. (C) Contribution to CID value by different sub-samples. C: control; S: shoot; R: root. The dashed horizontal line indicates the nominal value 1/26. (D) Marginal CDF (black solid line) and conditional CDF's under 0.5H_R, 0.5H_S (red dashed lines), 1H_R, 1H_S (green dashed lines), 12H_R, 12H_S (pink dashed lines), 24H_R, and 24H_S (yellow dashed lines).	28



3.5 Expression profiles and CID/pCID inferences of 253114\_at based on expression levels of *CBF1* and *CBF3*. (A) Averages and standard deviations of log<sub>2</sub> expression levels over time under control (CTRL) or cold treatments. (B) Contribution to CID(253114\_at|*CBF1*) and (C) pCID(253114\_at|*CBF3*; *CBF1*) by different sub-samples. C: control; S: shoot; R: root. The dashed horizontal line indicates the nominal value 1/26. . . . . 30

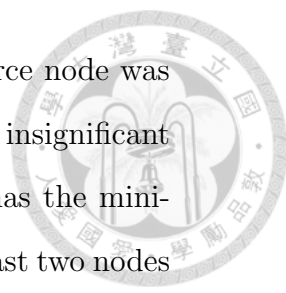
3.6 Number of the relevant variable  $X_i$  ( $i = 1, 2, 3, 5, 6$ ) being selected in 100 simulated samples of size (A) 100, (B) 50, or (C) 25 from the model  $Y = 10 \sin(\pi X_1 X_2) + 30(X_3 - 0.5)^2 + 10X_4 + 5X_5 + \epsilon$ , where  $X_i$ 's were distributed as  $U(0, 1)$  and  $\epsilon$  was distributed as  $N(0, 1)$ . . . 32

4.1 Diagram of gene regulatory network inference workflow. (A) Identification of a significantly associated gene pair. (B) Regulation path elongation. (C) Assembly of all identified regulation paths. . . . . 34

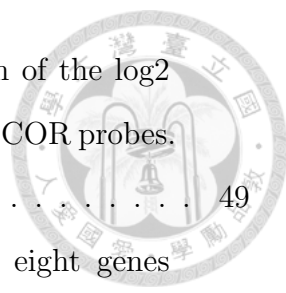
4.2 Illustration of the heuristic approach for regulation path elongation. . . 36

4.3 Illustration of the simple example for regulation path elongation used by CID/pCID method. . . . . 38

4.4 Pseudo network for the simulation study. The numbers next to the arrows illustrate the proportions of the objects in the sample that the expressions of the target node actually determined by the expressions of the source node. . . . . 40



- 4.5 The results of the network reconstructed under the source node was A11 based on the procedure in Section 4.1 (Exclude the insignificant node by CID, and pick up the connected node which has the minimum significant CID/pCID  $p$ -value, if there existed at least two nodes which fitted the requests, we chose the node that had the maximum CID/pCID value) from 100 simulations of pseudo network for  $N = 25, 50$  and  $100$ , respectively. The numbers next to the arrows illustrate the number of connection from the source node to the target node; besides, the number of connection in the brackets illustrated the inverse direction. . . . . 43
- 4.6 Pseudo network for the simulation study based on the procedure in Section 4.1 (Exclude the insignificant node by CID, and pick up the connected node which has the minimum significant CID/pCID  $p$ -value, if there existed at least two nodes which fitted the requests, we chose the node that had the maximum CID/pCID value). (A) The numbers next to the arrows illustrate the proportions of the objects in the sample that the expressions of the target node actually determined by the expressions of the source node. (B), (C) and (D) were the results which were combined with all connection from 100 simulations when the source node  $T_0$  was A11 for  $N = 25, 50$  and  $100$ , respectively. . . . . 44
- 4.7 Reconstruction of CBF-COR regulatory network with eight genes under cold stress was based on CID/pCID method. (A) Combination of the pathways from three source genes ( $CBF1$ ,  $CBF2$  and  $CBF3$ ). (B), (C) and (D) were the pathways from the source genes,  $CBF1$ ,  $CBF2$  and  $CBF3$ , respectively. Rectangle nodes indicate the source genes. Ellipse nodes are the candidate target genes. . . . . 48

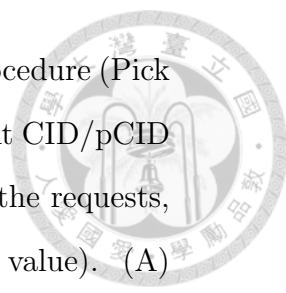


4.8 Cluster analysis and heatmap. A heatmap visualization of the log<sub>2</sub> relative treatment gene expression levels for the CBF and COR probes. R, root; S, shoot. . . . . 49

4.9 Reconstruction of CBF-COR regulatory network with eight genes under cold stress was based on CID/pCID method. (A) Combination of the pathways from all source genes (three CBF and five COR genes). (B), (C), (D), (E) and (F) were the pathways from the source genes, *COR47*, *COR6.6*, *COR78*, *COR15A* and *COR15B*, respectively. Rectangle nodes indicate the source genes. Ellipse nodes are the candidate target genes. . . . . 51

4.10 The gene regulatory network for OsbHLH rice seedlings contained the G-box binders and sequences under abiotic stresses is constructed by CID/pCID method from the NCBI-GEO database. Each node is the code of the OsbHLH number, for example 152 means the OsbHLH152. An arrow between nodes indicates a connection is determined by CID/pCID. Gray nodes show the genes are related to abiotic stresses have been confirmed from paper or GO term. Rectangle nodes indicate the OsbHLH probes are the G-box binders and exclude G-box sequences. Ellipse nodes indicate the OsbHLH probes include G-box sequences and are not the G-box binders. Octagon nodes are the G-box binders and include G-box sequences at the same time. . . . . 54

4.11 The results of the network reconstructed from 100 simulations of pseudo network for  $N = 25, 50$  and  $100$ , respectively. The numbers next to the arrows illustrate the number of connection from the source node to the target node; besides, the number of connection in the brackets illustrated the inverse direction. . . . . 57



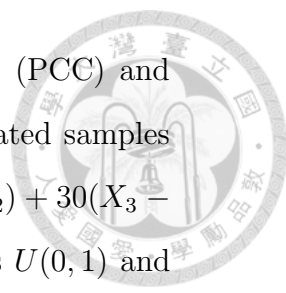
4.12 Pseudo network for the simulation study based on the procedure (Pick up the connected node which has the minimum significant CID/pCID  $p$ -value, if there existed at least two nodes which fitted the requests, we chose the node that had the maximum CID/pCID value). (A)

The numbers next to the arrows illustrate the proportions of the objects in the sample that the expressions of the target node actually determined by the expressions of the source node. (B), (C) and (D) were the results which were combined with all connection from 100 simulations when the source node  $T_0$  was A11 for  $N = 25, 50$  and 100, respectively. . . . . 58



# List of Tables

1.1	Summary statistics of univariate CID and bivariate CID values based on 100 simulated samples of size $N = 100$ from the model $Y = 10 \sin(\pi X_1 X_2) + 30(X_3 - 0.5)^2 + 10X_4 + 5X_5 + \varepsilon$ , where $X_i$ 's were distributed as $U(0, 1)$ and $\varepsilon$ was distributed as $N(0, 1)$ . . . . .	3
3.1	Summary statistics of univariate CID, bivariate CID, and pCID values based on 100 simulated samples of size $N = 100$ from the model $Y = 10 \sin(\pi X_1 X_2) + 30(X_3 - 0.5)^2 + 10X_4 + 5X_5 + \varepsilon$ , where $X_i$ 's were distributed as $U(0, 1)$ and $\varepsilon$ was distributed as $N(0, 1)$ . . . . .	17
3.2	Proportion (%) of negative pCID values based on 100 simulations from the model $Y = 10 \sin(\pi X_1 X_2) + 30(X_3 - 0.5)^2 + 10X_4 + 5X_5 + \varepsilon$ of samples size $N = 25, 50,$ and $100$ , where $X_i$ 's were distributed as $U(0, 1)$ and $\varepsilon$ was distributed as $N(0, 1)$ . . . . .	19
3.3	Summary statistics of CID and pCID values based 100 simulated samples of size 25, 50, or 100 from the model $Y = 10 \sin(\pi X_1 X_2) + 30(X_3 - 0.5)^2 + 10X_4 + 5X_5 + \varepsilon$ , where $X_i$ 's were distributed as $U(0, 1)$ and $\varepsilon$ was distributed as $N(0, 1)$ . The numbers in parenthese indicate the proportion of significant CID / pCID values at $\alpha = 0.05$ in 100 simulations. . . . .	20



3.4	Summary statistics of Pearson's correlation coefficients (PCC) and partial correlation coefficients (pPCC) based 100 simulated samples of size 25, 50, or 100 from the model $Y = 10 \sin(\pi X_1 X_2) + 30(X_3 - 0.5)^2 + 10X_4 + 5X_5 + \varepsilon$ , where $X_i$ 's were distributed as $U(0, 1)$ and $\varepsilon$ was distributed as $N(0, 1)$ . The numbers in parentheses indicate the proportion of significant CID / pCID values at $\alpha = 0.05$ in 100 simulations. . . . .	22
3.5	Information for 29 GO accessions identified as being significantly enriched according to CID/pCID significance. . . . .	26
4.1	The estimated CID and pCID values in one of the 100 simulations with sample size $N = 50$ . . . . .	41
4.2	Summary of the estimated CID/pCID values in 100 simulations with sample size $N = 25, 50$ and 100. . . . .	46
4.3	The estimated CID and pCID values in one of the 100 simulations with sample size $N = 50$ . . . . .	56
B.1	GenBank accession number of OsbHLH members is in this study. . .	68



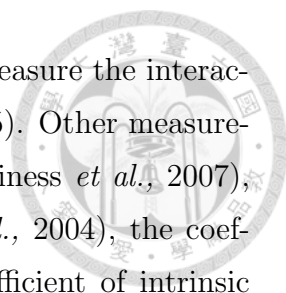


# Chapter 1

## Introduction

*Association* is defined as the correlation between explanatory and target variables. The type of variable involves discrete or continuous and the number of variables is univariate or multivariate. The association between two variables may exist linear, nonlinear or mixture relationship in reality. In this study, we explore the expressions of thousands of genes in biological microarray technology. One typical application is variable selection, feature selection in the words of machine learning, which used to identify the most relevant genes from thousands of gene expressions. These selected genes can provide some informations to biologists to verify an experiment further.

The other application can be extend to construct the gene regulatory network (GRN). Genes encode the information necessary for life which can be pass down the central dogma of molecular biology and translate proteins directly involving in different biological activities. Therefore, the expression level, or the amount of mRNA transcripts, partly reflects the activity of the gene. The gene expression levels of some genes are regulated by mRNAs of other genes or their protein products. This kind of gene regulation events can be possibly monitored using modern high-throughput gene expression technologies, including microarray or next generation sequencing (Mardis, 2008; Jain, 2012; Shrinet *et al.*, 2014). The gene regulation events under certain condition serve as small blocks to the entire gene regulation network (GRN), which may be reconstructed by connecting multiple regulation modules. An inferred GRN can therefore provide insights into the relationships between genes of interest by experiments and the understanding of biological functions with complex biological phenomena (Krouk *et al.*, 2013). More specifically, an inferred GRN consisting of the nodes (representing genes) and the edges (representing significant gene-gene interaction) reflects the gene regulation events that may concurrently or sequentially occur under the condition of study. In this study, we focus on the inference of GRN using the results of microarray experiments.



Pearson correlation coefficient (PCC) is mostly adopted to measure the interaction of genes based on their expression levels (Schadt *et al.*, 2005). Other measurements of association including the mutual information (MI) (Priness *et al.*, 2007), the partial Pearson correlation coefficient (pPCC) (Fuente *et al.*, 2004), the coefficient of determination (CoD) (Suh *et al.*, 2003), and the coefficient of intrinsic dependence (CID) (Hsing *et al.*, 2005; Liu, 2005; Liu *et al.*, 2009; Tsai and Liu, 2013) were also used. PCC and pPCC have the limitation of only identifying linear relationship between two gene expressions. In contrast, CID requires neither distributional (e.g. normal) nor functional (e.g. linear) assumptions on gene expression data. CID( $Y|X$ ) designates the CID value of a variable  $Y$  given the information of another variable  $X$ . It takes any real value between 0 and +1 inclusive. It is +1 in the case of full dependence and is 0 in the case of independence. As the level of dependence ascends, the CID value goes from 0 to 1. It was used to construct an estrogen receptor regulatory network in accompany with the correlation coefficient (Liu *et al.*, 2009), to infer and classify co-regulatory events by two transcription factors (Liu *et al.*, 2012), and to perform gene set association analysis (GSAA) (Tsai and Liu, 2013). We have demonstrated that CID outperformed the conventional methods in identification of different association patterns (Liu *et al.*, 2009; Tsai and Liu, 2013).

This study was initially motivated by the inquiry to select relevant explanatory variables to the target variable using CID. We used a toy example to illustrate the situation one might encountered when selecting variables using CID. Let  $Y$  be a one-dimensional target variable and  $X_i$ 's ( $i = 1, 2, \dots, 6$ ) be the one-dimensional candidate explanatory variables identically and independently distributed as Uniform(0, 1). In fact,

$$Y = 10 \sin(\pi X_1 X_2) + 30(X_3 - 0.5)^2 + 10X_4 + 5X_5 + \varepsilon, \quad (1.1)$$

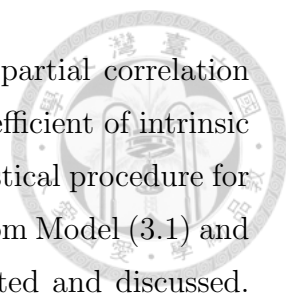
where  $\varepsilon$  is the random disturbance distributed as normal with zero mean and unit variance. Note that the explanatory variable  $X_6$  is independent of the target variable  $Y$  according to the model. Ideally, a proper stepwise procedure iteratively picks the relevant  $X_i$ 's according to its magnitude of association to  $Y$  until no more  $X_i$  would significantly increase the amount of association. Table 1.1 lists the summary statistics for the univariate CID values of  $Y$  given one of the explanatory variables and partially bivariate CID values based on 100 simulated samples of sizes  $N = 100$ . According to the result, CID( $Y|X_4$ ) had the largest value in average among all CID( $Y|X_i$ ) ( $i = 1, \dots, 6$ ) and was concluded as the most relevant predictors with  $Y$ .

Table 1.1: Summary statistics of univariate CID and bivariate CID values based on 100 simulated samples of size  $N = 100$  from the model  $Y = 10 \sin(\pi X_1 X_2) + 30(X_3 - 0.5)^2 + 10X_4 + 5X_5 + \varepsilon$ , where  $X_i$ 's were distributed as  $U(0, 1)$  and  $\varepsilon$  was distributed as  $N(0, 1)$ .

	Mean	SD	Proportion of Significant CID's		
			$\alpha = 0.1$	$\alpha = 0.05$	$\alpha = 0.01$
CID( $Y X_1$ )	0.0664	0.0238	0.99	0.98	0.95
CID( $Y X_2$ )	0.0683	0.0270	1.00	0.98	0.96
CID( $Y X_3$ )	0.0366	0.0142	0.93	0.83	0.65
CID( $Y X_4$ )	0.1176	0.0325	1.00	1.00	1.00
CID( $Y X_5$ )	0.0328	0.0202	0.74	0.69	0.45
CID( $Y X_6$ )	0.0077	0.0048	0.03	0.02	0.00
CID( $Y X_1, X_4$ )	0.1747	0.0319	1.00	1.00	1.00
CID( $Y X_2, X_4$ )	0.1783	0.0328	1.00	1.00	1.00
CID( $Y X_3, X_4$ )	0.1464	0.0279	1.00	1.00	1.00
CID( $Y X_5, X_4$ )	0.1415	0.0324	1.00	1.00	1.00
CID( $Y X_6, X_4$ )	0.1191	0.0309	1.00	1.00	0.99

To determine the second most relevant predictor, we further computed the bivariate CID values given  $X_4$  and another predictor  $X_i$ , CID( $Y|X_4, X_i$ ) ( $i = 1, 2, 3, 5, 6$ ) (Liu *et al.*, 2009). Due to the dominant influence from  $X_4$ , the two-predictor CID values were frequently claimed significant even if an irrelevant predictor, i.e.  $X_6$ , was added (Table 1.1). The above scenario was similar with the computation of regression coefficient,  $R^2$ , in a regression analysis – the more variables included in the model, the larger the CID value. This also implied a significant CID value of the target variable given multiple predictors once the most relevant variable was included although the other may not have strong association with the target.

The toy example implied the need of alternatives to evaluate the significance under stepwise variable selection to study the 'pure effect' coming from the variable of interest without disturbing by the other predictors. The process should also be able to justify different levels or types of association. Inspired by the partial correlation coefficient (pPCC), we proposed a new measure called partial coefficient of intrinsic dependence (pCID). The pPCC aims to describe the linear relationship of the target variable and the second predictor variable which cannot be explained by their respective linear relationship with the first predictor variable (Baba *et al.*, 2004). Similarly, pCID proposed in this study will further decompose the variability of distribution of the target variable which was not explained by the conditional distribution of the target variable given the first predictor.



In the next chapter, coefficient of intrinsic dependence and partial correlation coefficient will be reviewed and our proposed method, partial coefficient of intrinsic dependence, will be introduced. In Chapter 3, the proposed statistical procedure for stepwise variable selection will be given. The simulation design from Model (3.1) and compared results of CID/pCID and PCC/pPCC will be presented and discussed. A reality example using published microarray dataset in *Arabidopsis* illustrates the proposed method. In Chapter 4, the heuristic approach will be advanced to construct the gene regulatory network and will be used to reconstruct the pseudo network. The proposed procedure will practice on reconstruction cold-stress responsive regulation paths in *Arabidopsis* based on a microarray experiment and will provide an unverified gene network for biologists. The final conclusions are provided in Chapter 5.



## Chapter 2

# Partial Coefficient of intrinsic dependence (pCID)

In the current methods of association, coefficient of intrinsic dependence (CID) does not need common restrictions such as the type of variable and distributional or functional assumptions. Besides, CID had been demonstrated that have good performances in identification, classification, construction of gene regulatory network, performance of gene set association analysis (Liu, 2005; Liu *et al.*, 2009; Liu *et al.*, 2012; Tsai and Liu, 2013).

CID can find how much information of the target variable be explained by the predictor variables. Therefore, the CID value of the target variable is increasing as more predictor variables included. In Chapter 1, the toy example has been observed that the multivariate CID value was significant when the most relevant predictor variable was included even though the other irrelevant predictor variable was added. To solve this problem, we propose a new measure called partial coefficient of intrinsic dependence (pCID). Main objective in this study is to sift out the actual relevant predictors step by step. The concept of pCID is inspired by the partial Pearson correlation coefficient (pPCC). In this chapter, we describe the CID and pPCC in detail and introduce our method, pCID. And then we explain how to perform a hypothesis test of independence.

## 2.1 Coefficient of intrinsic dependence (CID)

Consider a pair of random variables  $(X, Y)$ , where  $X$  is a predictor variable and  $Y$  is a target variable. The general definition of the coefficient of intrinsic dependence,  $\text{CID}(Y|X)$ , is defined as follow (Liu, 2005):

$$\text{CID}(Y|X) = \frac{\int_{-\infty}^{\infty} \text{Var}_X\{E_{Y|X}[I(Y \leq u)]\}dF_Y(u)}{\int_{-\infty}^{\infty} \text{Var}_Y[I(Y \leq v)]dF_Y(v)}, \quad (2.1)$$

where  $F_Y(\cdot)$  is the marginal cumulative distribution function of  $Y$ , and  $I(\cdot)$  is an indicator function. If multiple predictors are considered, we let  $X = \{X_1, \dots, X_k\}$ , where  $k \geq 2$ . Then CID can be similarly defined (Tsai and Liu, 2013):

$$\text{CID}(Y|X_1, \dots, X_k) = \text{CID}(Y|\mathbf{X}) = \frac{\int_{-\infty}^{\infty} \text{Var}_{\mathbf{X}}\{E_{Y|\mathbf{X}}[I(Y \leq u)]\}dF_Y(u)}{\int_{-\infty}^{\infty} \text{Var}_Y[I(Y \leq v)]dF_Y(v)}, \quad (2.2)$$

The numerator of CID accounts the discrepancy between the marginal cumulative distribution function (cdf) of  $Y$  and the conditional cdf of  $Y$  given  $X$  as the amount of dependency between  $Y$  and  $X$ . The dependency (in the numerator) is then normalized between 0 and 1 by the denominator for the convenience of interpretation. If  $X$  and  $Y$  are nearly independent,  $X$  provides little information about  $Y$ . The independency causes the conditional and marginal distributions of  $Y$  similar to each other and the numerator of CID close 0. On the other hand, if  $X$  and  $Y$  are highly relevant, the information of  $X$  can almost surely predict the behavior of  $Y$ . In these cases, CID yields values close to 1.

It has been shown that the CID has several properties. CID can be carried out in different instances, such as all types of random variables (discrete, continuous, or including both ones) and multivariate cases. CID is a model-free measure in that it depends on calculating the estimator with a different sample. For that reason, CID does not require some common assumptions like normal and linear. CID is asymmetric, that is to say,  $\text{CID}(Y|X)$  does not remain the same as  $\text{CID}(X|Y)$ . Accordingly, CID takes the causal relationship between variables into account.

## 2.2 Partial coefficient of intrinsic dependence (pCID)

Inspired by the partial correlation coefficient, the coefficient of partial coefficient of intrinsic dependence (pCID) further decomposes the variability of distribution of the target variable. Let  $Y$  be the target variable,  $X_1$  be the first dominant predictor variable, and  $X_2$  be the second dominant predictor variable. By definition, if  $Y$  and  $X_2$  are independent given the values of  $X_1$  if and only if

$$F(y, x_2|x_1) = F(x_2|x_1)F(y|x_1),$$

and

$$\begin{aligned} F(y|x_1, x_2) &= \frac{F(x_1, x_2, y)}{F(x_1, x_2)} = \frac{F(y, x_2|x_1)F(x_1)}{F(x_1, x_2)} = \frac{F(x_2|x_1)F(y|x_1)F(x_1)}{F(x_1, x_2)} \\ &= \frac{[F(x_1, x_2)/F(x_1)][F(x_1, y)/F(x_1)]F(x_1)}{F(x_1, x_2)} = \frac{F(x_1, y)}{F(x_1)} = F(y|x_1), \end{aligned}$$

where  $F$ 's are corresponding conditional or marginal cumulative distribution functions. Hence, the discrepancy between two conditional distributions  $F(y|x_1, x_2)$  and  $F(y|x_1)$  represents the amount of dependency between  $Y$  and  $X_2$  given  $X_1$ . The Cramér-von Mises distance between the two distributions can be expressed as

$$\int_{-\infty}^{\infty} \{F(y|x_1, x_2) - F(y|x_1)\}^2 dF_Y(y). \quad (2.3)$$

To average out the different values of  $x_1$ 's and  $x_2$ 's, we take expectations over  $X_1$  and  $X_2$ , respectively. The expectations over  $X_1$  and  $X_2$  were taken to average out the effects from different values of  $x_1$ 's and  $x_2$ 's. Hence, Equation (2.3) can be revised as follow:

$$\begin{aligned} & \int_{-\infty}^{\infty} \mathbb{E}_{X_1} \mathbb{E}_{X_2} \{F(y|x_1, x_2) - F(y|x_1)\}^2 dF_Y(y) \\ &= \int_{-\infty}^{\infty} \mathbb{E}_{X_1} \mathbb{E}_{X_2} \{P(Y \leq y|x_1, x_2) - P(Y \leq y|x_1)\}^2 dF_Y(y) \\ &= \int_{-\infty}^{\infty} \mathbb{E}_{X_1} \mathbb{E}_{X_2} \{\mathbb{E}_{Y|x_1, x_2}[I(Y \leq y)] - \mathbb{E}_{Y|x_1}[I(Y \leq y)]\}^2 dF_Y(y) \\ &= \int_{-\infty}^{\infty} \mathbb{E}_{X_1} \text{Var}_{X_2} \{\mathbb{E}_{Y|x_1, x_2}[I(Y \leq y)]\} dF_Y(y), \end{aligned} \quad (2.4)$$

where  $I(\cdot)$  is an indicator function. The coefficient of partial intrinsic dependence of  $Y$  given  $X_2$  conditioned on  $X_1$  was defined by standardized Equation (2.4) using variance decomposition:

$$\text{pCID}(Y|X_2; X_1) = \frac{\int_{-\infty}^{\infty} \mathbb{E}_{X_1} \text{Var}_{X_2} \{\mathbb{E}_{Y|x_1, x_2}[I(Y \leq u)]\} dF_Y(u)}{\int_{-\infty}^{\infty} \mathbb{E}_{X_1} \text{Var}_{Y|X_1} [I(Y \leq v)] dF_Y(v)}. \quad (2.5)$$

Given the target variable takes distinct values on a continuous domain, the denominator of  $\text{pCID}(Y|X_2; X_1)$  can be expressed as

$$\begin{aligned} \int_{-\infty}^{\infty} \mathbb{E}_{X_1} \text{Var}_{Y|X_1} [I(Y \leq v)] dF_Y(v) &= \int_0^1 \mathbb{E}_{X_1} \text{Var}_{Y|X_1} [I(F_Y(Y) \leq v)] dv \\ &= \int_0^1 \mathbb{E}_{X_1} \{\mathbb{E}_{Y|x_1} [I^2(F_Y(Y) \leq v)] - [\mathbb{E}_{Y|x_1} [I(F_Y(Y) \leq v)]]^2\} dv \\ &= \int_0^1 \mathbb{E}_Y [I(F_Y(Y) \leq v)] - \mathbb{E}_{X_1} [\mathbb{E}_{Y|x_1} [I(F_Y(Y) \leq v)]]^2 dv \\ &= \int_0^1 v dv - \int_0^1 \mathbb{E}_{X_1} [P^2(F_Y(Y) \leq v)|x_1] dv \\ &= \frac{1}{2} - \int_0^1 \mathbb{E}_{X_1} [P^2(Y \leq F_Y^{-1}(v))|x_1] dv \end{aligned}$$



Similarly, the numerator of  $\text{pCID}(Y|X_2; X_1)$  is

$$\begin{aligned}
& \int_{-\infty}^{\infty} \mathbb{E}_{X_1} \text{Var}_{X_2} \{ \mathbb{E}_{Y|x_1, x_2} [I(Y \leq u)] \} dF_Y(u) \\
&= \int_0^1 \mathbb{E}_{X_1} \text{Var}_{X_2} \{ \mathbb{E}_{Y|x_1, x_2} [I(F_Y(Y) \leq u)] \} du \\
&= \int_0^1 \mathbb{E}_{X_1} \{ \mathbb{E}_{X_2} [ [\mathbb{E}_{Y|x_1, x_2} [I(F_Y(Y) \leq u)]^2] - [\mathbb{E}_{X_2} [\mathbb{E}_{Y|x_1, x_2} [I(F_Y(Y) \leq u)]]]^2 ] \} du \\
&= \int_0^1 \mathbb{E}_{X_1} \{ \mathbb{E}_{X_2} [P^2(F_Y(Y) \leq u|x_1, x_2)] - P^2(F_Y(Y) \leq u|x_1) \} du \\
&= \int_0^1 \mathbb{E}_{X_1} \mathbb{E}_{X_2} [P^2(Y \leq F_Y^{-1}(u)|x_1, x_2)] du - \int_0^1 \mathbb{E}_{X_1} [P^2(Y \leq F_Y^{-1}(u)|x_1)] du
\end{aligned}$$

Hence, for the continuous target variable  $Y$ ,

$$\text{pCID}(Y|X_2; X_1) = \frac{\int_0^1 \mathbb{E}_{X_1} \mathbb{E}_{X_2} [P^2(Y \leq F_Y^{-1}(u)|x_1, x_2)] du - \int_0^1 \mathbb{E}_{X_1} [P^2(Y \leq F_Y^{-1}(u)|x_1)] du}{\frac{1}{2} - \int_0^1 \mathbb{E}_{X_1} [P^2(Y \leq F_Y^{-1}(v))|x_1] dv}.$$

According to the CID formula for the continuous target (Liu, 2005),

$$\text{CID}(Y|X) = 6 \int_0^1 \mathbb{E}_X [P^2(Y \leq F_Y^{-1}(y))|x] dy - 2,$$

the following recursive formula can be derived to compute the coefficient of partial intrinsic dependence of  $Y$  given  $X_2$  conditioned on  $X_1$ :

$$\begin{aligned}
\text{pCID}(Y|X_2; X_1) &= \frac{\frac{1}{6}[\text{CID}(Y|X_1, X_2) + 2] - \frac{1}{6}[\text{CID}(Y|X_1) + 2]}{\frac{1}{2} - \frac{1}{6}[\text{CID}(Y|X_1) + 2]} \\
&= \frac{\text{CID}(Y|X_1, X_2) - \text{CID}(Y|X_1)}{1 - \text{CID}(Y|X_1)}, \tag{2.6}
\end{aligned}$$

where  $\text{CID}(Y|X_1, X_2)$  and  $\text{CID}(Y|X_1)$  are the ordinary coefficients of intrinsic dependence of  $Y$  given  $X_1, X_2$  and of  $Y$  given  $X_1$ , respectively. Similarly,  $\text{pCID}$  takes any real values between 0 and +1 inclusive; it is +1 in the case of full dependence between  $Y$  and  $X_2$  given the value of  $X_1$  and is zero in the case of independence. As the level of dependence ascends, the value of  $\text{pCID}$  goes from 0 to 1.  $\text{pCID}(Y|X_2; X_1)$  can be estimated from data by using the recursive formula and plugging in the corresponding estimated CID values. Similarly, the coefficient of partial intrinsic dependence of  $Y$  given  $X_i$  conditioned on  $\{X_1, X_2, \dots, X_{i-1}\}$  can be derived as

$$\text{pCID}(Y|X_i; \{X_1, \dots, X_{i-1}\}) = \frac{\text{CID}(Y|X_1, \dots, X_i) - \text{CID}(Y|X_1, \dots, X_{i-1})}{1 - \text{CID}(Y|X_1, \dots, X_{i-1})}.$$

## 2.3 Estimation of CID and pCID

According to the definition of CID is not under any assumption, the marginal and conditional distributions have to be estimated from the sample by the empirical



distribution function. In section 2.1, CID is defined separately by unitary and multiple predictors. Let  $(x_i, y_i)$  be the  $i$ th paired observation of the random variables  $(X, Y)$  from a sample size of  $N$ , where  $i = 1, \dots, N$ . The estimator of CID (Equation (2.1)) is

$$\text{CID}(Y|X) = \frac{1}{N} \times \frac{\sum_{i=1}^N \sum_{j=1}^N \left[ \hat{F}(y_i|x_j) - \hat{F}(y_i) \right]^2}{\sum_{i=1}^N \hat{F}(y_i) \left[ 1 - \hat{F}(y_i) \right]},$$

where  $x_j$  be the observed value of  $X$  in the  $j$ th object. If  $\mathbf{X}$  is  $k$ -dimensional predictor variable ( $k \geq 2$ ),  $\mathbf{x}_j$  be the vector containing observations of  $\{X_1, \dots, X_k\}$  in the  $j$ th object. Then the estimated value of CID (Equation 2.2) is as follows:

$$\text{CID}(Y|X_1, \dots, X_k) = \text{CID}(Y|\mathbf{X}) = \frac{1}{N} \times \frac{\sum_{i=1}^N \sum_{j=1}^N \left[ \hat{F}(y_i|\mathbf{x}_j) - \hat{F}(y_i) \right]^2}{\sum_{i=1}^N \hat{F}(y_i) \left[ 1 - \hat{F}(y_i) \right]}. \quad (2.7)$$

In previous studies (Liu, 2005; Liu *et al.*, 2009; Liu *et al.*, 2012; Tsai and Liu, 2013), the estimate of CID relies on subgrouping the sample of predictors  $\mathbf{X}$  to calculate the value of conditional distribution function,  $\hat{F}(y|\mathbf{x})$ . The subgroup is used to place the sample of size  $N$  into  $P$  subgroups according to the observed values of  $\mathbf{X}$ . In each subgroup  $s$  ( $s = 1, \dots, P$ ), the estimate of the cumulative marginal and conditional distribution functions are below.

$$\begin{aligned} \hat{F}(y_i) &= \frac{1}{N} \sum_{q=1}^N I(y_q < y_i), \\ \hat{F}_s(y_i) &= \frac{1}{N_s} \sum_{q=1}^N I(y_q < y_i \text{ and } \mathbf{x}_q \in \text{the } s\text{th subgroup}), \\ \text{and } N_s &= \sum_{j=1}^N I(\mathbf{x}_j \in \text{the } s\text{th subgroup}) \end{aligned}$$

A weighted average is taken to account all discrepancies measured within different subgroups and yields the estimate of CID:

$$\text{CID}(Y|\mathbf{X}) = \frac{\sum_{i=1}^N \sum_{s=1}^P \frac{N_s}{N} \left[ \hat{F}_s(y_i) - \hat{F}(y_i) \right]^2}{\sum_{i=1}^N \hat{F}(y_i) \left[ 1 - \hat{F}(y_i) \right]}.$$

Two general sample subgrouping method, quantile and hierarchical clustering method, have been used commonly. The quantile method categorizes the  $m$ th dimension of  $\mathbf{X}$  into  $r_m$  subgroups with an equal or approximate equal number of observations in each subgroup. If  $\mathbf{X}$  has  $k$  dimensions, the sample is separated into  $P = \prod_{m=1}^k r_m$  subgroups. In general, the number of subgroups is set  $r_m = r$  for all

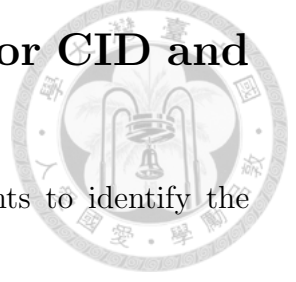
$m$  and  $P = r^k$  to fairly weight all dimensions of  $\mathbf{X}$ . However, it is in a predicament when  $k$  increases. This situation causes that the observations distribute sparsely and each subgroup has zero or too few observations. Besides, the quantile method has another problem, the number of subgroups is restricted. The hierarchical clustering method assigns a set of objects into  $P$  subgroups such that the objects in the same subgroup are more similar to each other. The result of the subgroup in the  $m$ th dimension of  $\mathbf{X}$  was changed when adding another predictor. This situation does not cause a problem in the estimated value of CID, but it influences the accuracy of the estimated value of pCID.

In this study, we propose the nonparametric kernel smoothing method using the 'np' package in R (version 0.40-13) (Hayfield and Racine, 2008) to estimate the corresponding distribution functions as follows.

$$\begin{aligned}\hat{F}(y_i) &= \int_{-\infty}^{y_i} \frac{1}{N} \sum_{q=1}^N \left[ \frac{K_Y\left(\frac{t-y_q}{h_Y}\right)}{h_Y} \right] dt \\ \text{and } \hat{F}(y_i | \mathbf{x}_j) &= \int_{-\infty}^{y_i} \frac{\frac{1}{N} \sum_{q=1}^N \left\{ \left[ \frac{K_Y\left(\frac{t-y_q}{h_Y}\right)}{h_Y} \right] \cdot \prod_{p=1}^k \left[ \frac{K_X\left(\frac{x_{pj}-x_{pq}}{h_p}\right)}{h_p} \right] \right\}}{\frac{1}{N} \sum_{q=1}^N \prod_{p=1}^k \left[ \frac{K_X\left(\frac{x_{pj}-x_{pq}}{h_p}\right)}{h_p} \right]} dt \\ &= \int_{-\infty}^{y_i} \frac{\sum_{q=1}^N \left\{ \left[ \frac{K_Y\left(\frac{t-y_q}{h_Y}\right)}{h_Y} \right] \cdot \prod_{p=1}^k K_X\left(\frac{x_{pj}-x_{pq}}{h_p}\right) \right\}}{\sum_{q=1}^N \prod_{p=1}^k K_X\left(\frac{x_{pj}-x_{pq}}{h_p}\right)} dt,\end{aligned}$$

where  $K(\cdot)$  is the kernel function with bandwidth  $h$ . We chose Second-Order Gaussian kernel,  $K(z) = \frac{\exp\left(-\frac{z^2}{2}\right)}{\sqrt{2\pi}}$ , for smoothing and the rule-of-thumb method for bandwidth selection. The formula of the rule-of-thumb bandwidth is  $h = 1.06\sigma N^{-\frac{1}{5}}$ , where  $\sigma$  is defined as the minimum value of measures of scale which are standard deviation (SD), mean absolute deviation (MAD)/1.4826 and interquartile range (IQR)/1.349. This method could solve the problems which the subgrouping methods produce. Therefore, the estimated values of CID and pCID are using the nonparametric kernel smoothing method to apply to simulations and real data studies.

## 2.4 Hypothesis test of Independence for CID and pCID



The hypothesis test for coefficient of intrinsic dependence points to identify the association between two samples as follows.

$$H_0 : Y \text{ does not depend on } X$$

$$H_1 : Y \text{ depends on } X$$

The null distribution of  $CID(Y|X)$  is difficult to formulate under assumption are ignored and will be generated by random permutations. We can chose the observed values of  $X$  or  $Y$  to be permuted randomly and the other values of variable are fixed. After that using these new combination in each run of random permutation to compute the CID value using Equation (2.7).

The partial coefficient of intrinsic dependence aims to test which of the following null and alternative hypotheses are preferred by observing the data:

$$H_0 : Y \text{ does not depend on } X_j, \text{ conditioned on } X_i$$

$$H_1 : Y \text{ depends on } X_j, \text{ conditioned on } X_i$$

Similarly, the null distribution of  $pCID(Y|X_j; X_i)$  will be generated by random permutations. But the selection of variable about random permutation would be changed. To keep the dependence between  $X_i$  and  $Y$ , only the values of  $X_j$  are randomly permuted. In other words, when we compute the  $pCID(Y|X_j; X_i)$  value from each run of random permutation,  $CID(Y|X_j, X_i)$  would be altered where the values of  $X_j$  are from permutation and  $CID(Y|X_i)$  are computed from the sample.

Random permutation was repeated  $R$  times and yielded  $R$  internal control values for each measure under independence. Let  $E_0$  be the estimate of an  $CID(Y|X)$  or  $pCID(Y|X_j; X_i)$  from the sample, and  $E_r$  be the estimate for that measure from the  $r$ th random permutation. The permuted  $p$ -value for  $CID(Y|X)$  or  $pCID(Y|X_j; X_i)$  was determined by

$$\frac{1}{R+1} \left( 1 + \sum_{r=1}^R I(E_r \geq E_0) \right). \quad (2.8)$$

## 2.5 The partial Pearson correlation coefficient (pPCC)

We compared the results of the partial coefficient of intrinsic dependence with that of the well-known partial correlation coefficient (pPCC). The partial correlation

coefficient describes the relationship between two variables after taking away the effect of another variable, or several other variables, on this relationship. The pPCC of  $Y$  and  $X_j$  adjusted for  $X_i$  is:

$$\text{pPCC}(Y, X_j; X_i) = \frac{r_{Y,X_j} - r_{Y,X_i}r_{X_j,X_i}}{\sqrt{(1 - r_{Y,X_i}^2)(1 - r_{X_j,X_i}^2)}},$$

where  $r_{U,V}$  is the Pearson's correlation coefficient (PCC) between two random variables  $U$  and  $V$ . The pPCC of  $Y$  with  $X_i$  given  $\{X_1, X_2, \dots, X_{i-2}, X_{i-1}\} = \{\mathbf{X}_{i-2}, X_{i-1}\}$  can be derived recursively:

$$\begin{aligned} & \text{pPCC}(Y, X_i; X_1, \dots, X_{i-1}) \\ &= \text{pPCC}(Y, X_i; \mathbf{X}_{i-2}, X_{i-1}) \\ &= \frac{\text{pPCC}(Y, X_i; \mathbf{X}_{i-2}) - \text{pPCC}(Y, X_{i-1}; \mathbf{X}_{i-2})\text{pPCC}(X_i, X_{i-1}; \mathbf{X}_{i-2})}{\sqrt{(1 - \text{pPCC}(Y, X_{i-1}; \mathbf{X}_{i-2})^2)(1 - \text{pPCC}(X_i, X_{i-1}; \mathbf{X}_{i-2})^2)}}. \end{aligned}$$

In most cases, the pPCC between two variables while removing the effect of the third variable is smaller than the PCC. But in the other cases where the absolute value of the pPCC becomes larger, the third variable may be a suppressor variable which can improve the association with two variables, but that is unrelated to the target variable. In this study, the pPCC value was calculated using the 'ppcor' package (version 1.0) in R (Kim, 2012).

A  $t$ -test statistic with  $N - 2 - k$  degrees of freedom, where  $k$  is the number of the controlling variables, can be yielded to access the significance of the partial correlation. However, in order to compare with our proposed method on the same basis, the  $p$ -values of the partial correlation will be obtained through  $R$  times of random permutation in this study similar with those of pCID (Equation 2.8).





## Chapter 3

# Application to stepwise variable selection

Variable selection, also known as feature selection, is the technique of picking up the relevant predictor variables with the target variable. In biometric, variable selection is ordinarily applied in microarray data which contains thousands of genes and a few tens to hundreds of samples. In order to explain the data more accurately, the redundant genes should be removed without resulting in much loss of data information. Further, stepwise variable selection is the process of selecting predictor variables step by step without the interference from other effect of variables. In this chapter, we construct the procedure of stepwise variable selection by using pCID and pPCC methods. Apply the procedure to simulation study and microarray data, and then compare the result of these methods.

### 3.1 The procedure for selecting variables

Forward selection is an approach of adding one variable which have the largest relationship at a time until none of remaining variables provides the statistical significance. According to this concept, we could find the important predictors with a target variable in order by pCID. The decision process by calculating pCID value is described below (see also Figure 3.1 ).

Suppose there are one target variable  $Y$  and  $k$  predictor variables  $\mathbf{X} = X_1, \dots, X_k$  from the sample size of  $N$ . First, calculate the all CID values of  $Y$  given each  $X_i$ , where  $i = 1, \dots, k$ , and then choose the most important predictor  $X_{(1)}$  which has the maximum value of  $\text{CID}(Y|X_i)$ . To get the  $p$ -value of  $\text{CID}(Y|X_{(1)})$ , we randomly permute  $X_{(1)}$  with  $R$  replicates. If the  $p$ -value of  $\text{CID}(Y|X_{(1)})$  was more than the significance level  $\alpha$ , the process would be ended. No predictor variables relate to this

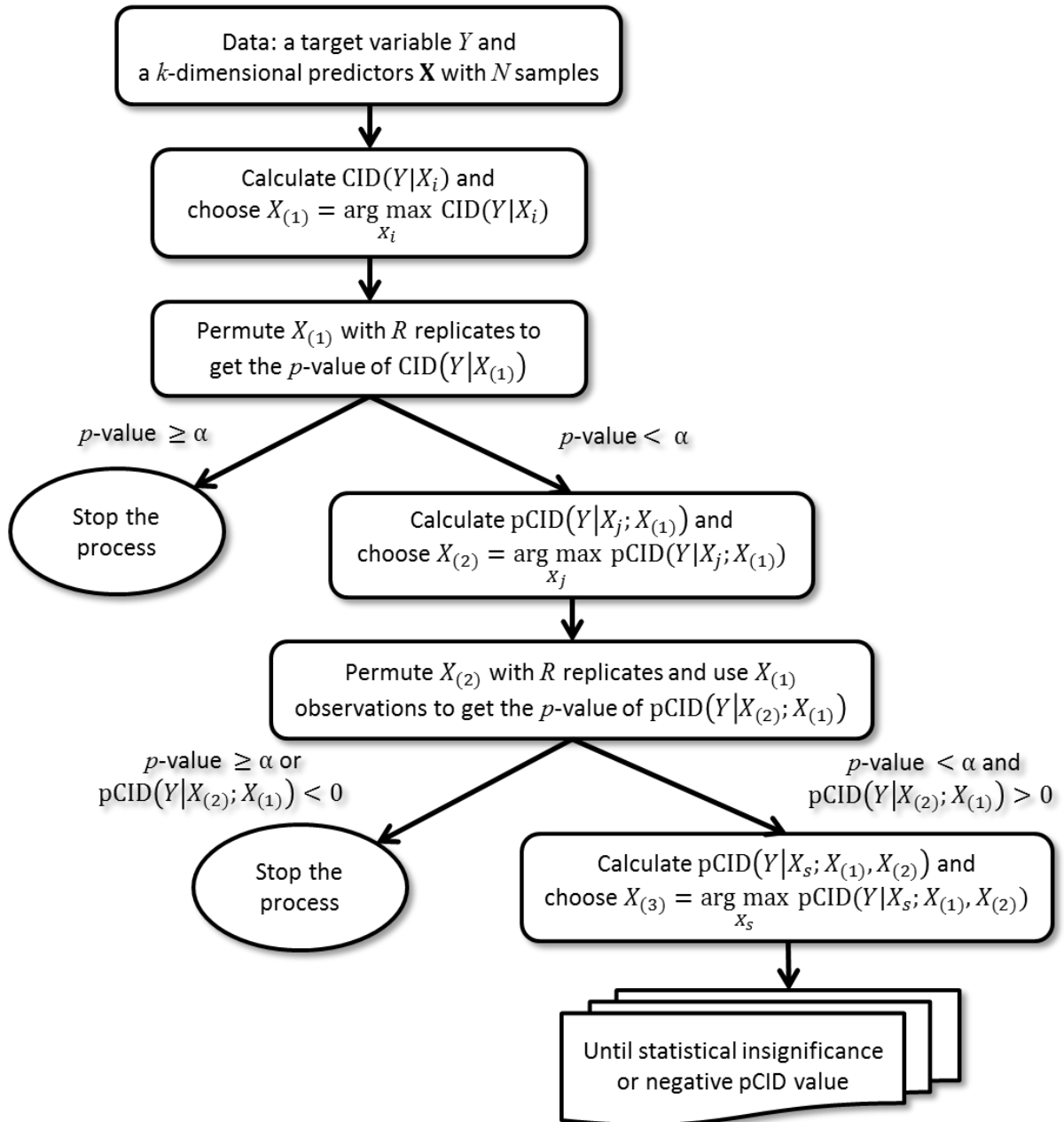


Figure 3.1: Flow chart of stepwise variable selection based on the CID and pCID.

target variable. Otherwise, the process proceeded and then calculate all pCID values of  $Y$  given each  $X_j$  conditioned on  $X_{(1)}$ , where  $j = 1, \dots, k - 1$  and  $X_j$  excluded  $X_{(1)}$ , to get the second important predictor  $X_{(2)}$  which has the maximum value of  $\text{pCID}(Y|X_j; X_{(1)})$ . The  $p$ -value of  $\text{pCID}(Y|X_{(2)}; X_{(1)})$  was calculated from the  $X_{(2)}$  permutation values and  $X_{(1)}$  observation values. Similarly, the process would be ended if  $\text{pCID}(Y|X_{(2)}; X_{(1)})$  was insignificantly dependent or negative, if not, the process still go forward to calculate all pCID values of  $Y$  given  $X_s$ , which is one of the other  $k - 2$  predictors, conditioned on  $X_{(1)}$  and  $X_{(2)}$ . The procedure for selecting variables was finished until the picked pCID value was insignificantly dependent or negative. Accordingly, the Pearson correlation coefficient (PCC) and the partial Pearson correlation coefficient (pPCC) can completely imitate this process to select relevant predictor variables.

## 3.2 Simulation study

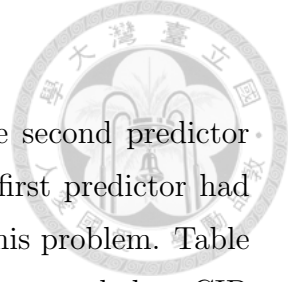
Our objective in variable selection is applying to pick up most relevant genes from thousands of gene expressions. Consider the relationship between two genes is not only linearity, we referred to the Friedman model (Friedman, 1991) and modified it as follows.

Suppose  $X_i$ 's ( $i = 1, \dots, 6$ ) were independent and identically distributed (*i.i.d.*) as  $Uniform(0, 1)$  and  $Y$  was determined by the following equation:

$$Y = 10 \sin(\pi X_1 X_2) + 30(X_3 - 0.5)^2 + 10X_4 + 5X_5 + \varepsilon, \quad (3.1)$$

where  $\varepsilon$  was distributed as  $Normal(0, 1)$ . In Model (3.1),  $X_1$  to  $X_5$  are dependent to  $Y$  while  $X_6$  is not.

The Pearson correlation coefficient (PCC) and the partial Pearson correlation coefficient (pPCC) are principal methods to discuss the relation of gene expressions in biological studies. We compared the results of the Partial coefficient of intrinsic dependence (pCID) to those of the pPCC in simulations of Model (3.1). Besides, we want to observe the effect upon the different sample size to generate a sample of size  $N$  ( $N = 25, 50, 100$ ). Then we consulted the procedure of variable selection which is detailed in Section 3.1 where a parameter  $k$  is equal to six. The simulation results of CID/pCID and PCC/pPCC are displayed in subsection 3.2.1 and 3.2.2, respectively.



### 3.2.1 The results of CID and pCID

As described in Section 1, bivariate CID could not identify the second predictor variable which associated with the target variable  $Y$  when the first predictor had strong relation with  $Y$ . We propose the pCID method to solve this problem. Table 3.1 presents the CID values of  $Y$  given either one or two predictors and the pCID values from 100 simulations and the sample of size  $N = 100$ . CID( $Y|X_4$ ) had the largest average value of CID( $Y|X_i$ ), 0.1176, for all  $i = 1, \dots, 6$ , meaning the distribution of  $Y$  was notably altered after conditioning on the values of  $X_4$ . A hundred  $p$ -values of CID( $Y|X_4$ ) were obtained from permuting the values of  $X_4$  with 1000 replicates. In Table 3.1, the proportions of significant CID( $Y|X_4$ ) at three different significant levels ( $\alpha = 0.1, 0.05, 0.01$ ) were 100%, which means all  $p$ -values of CID( $Y|X_4$ ) were smaller than 0.01. The variable associated with the target variable  $Y$  next to  $X_4$  in Model (3.1) was not selected based on the CID( $Y|X_i, X_4$ ) values for  $i = \{1, 2, 3, 5, 6\}$  but selected based on the pCID( $Y|X_i; X_4$ ) values for  $i = \{1, 2, 3, 5, 6\}$ . The proportions of significant CID( $Y|X_6, X_4$ ) were almost 100% and CID( $Y|X_6, X_4$ ) had large average value, 0.1191, even if  $X_6$  was not dependent on  $Y$  in Model (3.1). Besides, we observe the pCID( $Y|X_i; X_4$ ) values for  $i = \{1, 2, 3, 5, 6\}$  from different sample of sizes  $N = 25, 50$  and 100 by the boxplots which are presented in Figure 3.2. The variance of the pCID estimates would increase along with the increment of average pCID values. A relatively large sample size was necessary to obtain a consistent pCID estimate but the hypotheses testing of independence would already be quite effective under moderate sample size. According to the results of pCID( $Y|X_i; X_4$ ) values for  $i = \{1, 2, 3, 5, 6\}$ ,  $X_1, X_2$  were the most influential variables next to  $X_4$  toward  $Y$  by having the larger pCID values given  $X_4$ , while the random noise,  $X_6$ , had pCID( $Y|X_6; X_4$ ) closest to 0. The results of hypotheses testing for pCID( $Y|X_i; X_4$ )'s in Table 3.1, pCID( $Y|X_1; X_4$ ) and pCID( $Y|X_2; X_4$ ) had the largest average values (0.0644 and 0.0684, respectively) and more than 97% of the 100 pCID values were significant. The percentage of significant pCID( $Y|X_6; X_4$ ) values for irrelevant  $X_6$  were roughly consistent with the nominal significance levels and the average pCID( $Y|X_6; X_4$ ) value, 0.0015, was close to 0.

Sometimes pCID( $Y|X_i; X_4$ ) estimates had negative values (i.e., values below the grey horizontal line in Figure 3.2) which were not in the range of pCID values according to the definition. This might be due to the biased nature of the CID estimates, especially when the sample size is small (Liu, 2005). The pCID would inherit the bias if it was estimated using the recursive formula (i.e., Equation (2.6)).





Table 3.1: Summary statistics of univariate CID, bivariate CID, and pCID values based on 100 simulated samples of size  $N = 100$  from the model  $Y = 10 \sin(\pi X_1 X_2) + 30(X_3 - 0.5)^2 + 10X_4 + 5X_5 + \varepsilon$ , where  $X_i$ 's were distributed as  $U(0, 1)$  and  $\varepsilon$  was distributed as  $N(0, 1)$ .

	Mean	SD	Proportion of Significant CID's		
			$\alpha = 0.1$	$\alpha = 0.05$	$\alpha = 0.01$
CID( $Y X_1$ )	0.0664	0.0238	0.99	0.98	0.95
CID( $Y X_2$ )	0.0683	0.0270	1.00	0.98	0.96
CID( $Y X_3$ )	0.0366	0.0142	0.93	0.83	0.65
CID( $Y X_4$ )	0.1176	0.0325	1.00	1.00	1.00
CID( $Y X_5$ )	0.0328	0.0202	0.74	0.69	0.45
CID( $Y X_6$ )	0.0077	0.0048	0.03	0.02	0.00
CID( $Y X_1, X_4$ )	0.1747	0.0319	1.00	1.00	1.00
CID( $Y X_2, X_4$ )	0.1783	0.0328	1.00	1.00	1.00
CID( $Y X_3, X_4$ )	0.1464	0.0279	1.00	1.00	1.00
CID( $Y X_5, X_4$ )	0.1415	0.0324	1.00	1.00	1.00
CID( $Y X_6, X_4$ )	0.1191	0.0309	1.00	1.00	0.99
pCID( $Y X_1; X_4$ )	0.0644	0.0221	0.99	0.97	0.97
pCID( $Y X_2; X_4$ )	0.0684	0.0251	1.00	0.99	0.97
pCID( $Y X_3; X_4$ )	0.0322	0.0157	0.91	0.82	0.63
pCID( $Y X_5; X_4$ )	0.0268	0.0193	0.73	0.65	0.43
pCID( $Y X_6; X_4$ )	0.0015	0.0084	0.10	0.05	0.01

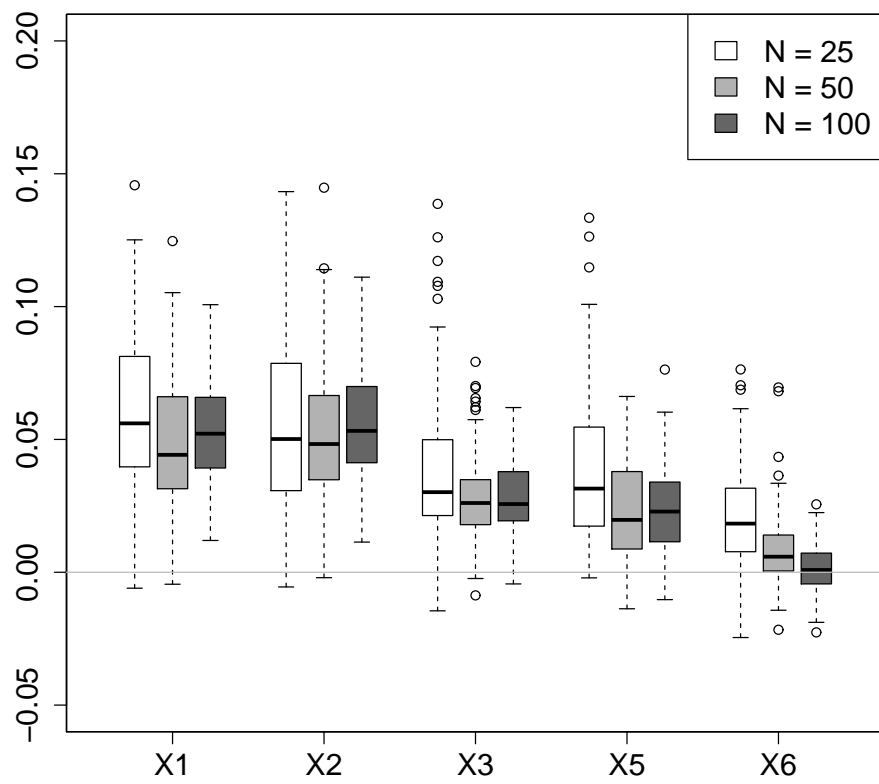
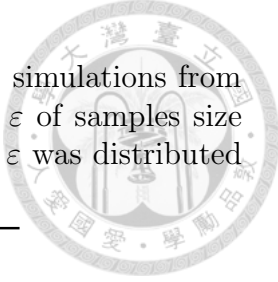


Figure 3.2: Boxplots of  $\text{pCID}(Y|X_i; X_4)$  values,  $i = 1, 2, 3, 5, 6$ , based 100 simulated samples of size 25, 50, or 100 from the model  $Y = 10 \sin(\pi X_1 X_2) + 30(X_3 - 0.5)^2 + 10X_4 + 5X_5 + \varepsilon$ , where  $X_i$ 's were distributed as  $U(0, 1)$  and  $\varepsilon$  was distributed as  $N(0, 1)$ . The horizontal line indicates the zero value.

Table 3.2: Proportion (%) of negative pCID values based on 100 simulations from the model  $Y = 10 \sin(\pi X_1 X_2) + 30(X_3 - 0.5)^2 + 10X_4 + 5X_5 + \varepsilon$  of samples size  $N = 25, 50,$  and  $100,$  where  $X_i$ 's were distributed as  $U(0, 1)$  and  $\varepsilon$  was distributed as  $N(0, 1)$ .



	Explanatory Variable					
	$X_1$	$X_2$	$X_3$	$X_4$	$X_5$	$X_6$
$N = 25$	1.95	3.73	2.97	0.00	2.42	5.70
$N = 50$	1.14	0.68	0.69	0.00	3.21	7.24
$N = 100$	0.00	0.00	0.91	0.00	2.33	14.44

The proportions of negative pCID values (Table 3.2) were less than 4% for the relevant variables (i.e.,  $X_1$  to  $X_5$ ), but the problem was elevated for the irrelevant variable  $X_6$ . Generally speaking, more negative values would be yielded when the average pCID value is closer to zero, and all the negative values were indeed close to 0 (the minimal negative pCID value was -0.022 in the entire simulation, and 84% of the negative values were greater than -0.01). These negative values can be avoided by using a larger sample size or using Equation (2.5) and directly estimating the corresponding conditional distributions.

Based on similar philosophy, the relevant variables can be consecutively selected according the corresponding CID/pCID values in a real practice. The summary statistics for sequentially selected CID/pCID values from all 100 simulations for samples of size  $N = 25, 50,$  and  $100$  are provided in Table 3.3. According to the average values of pCID, the order of the explanatory variables according to their importance toward  $Y$  is  $X_4, X_2, X_1, X_3,$  and  $X_5,$  while  $X_6$  was identified as being irrelevant to  $Y$ . Note that both the CID and pCID identified the same order of importance for the six explanatory variables regardless of the sample size. But the pCID controlled the type I error a bit better than the CID (Tables 3.1 and 3.3).



Table 3.3: Summary statistics of CID and pCID values based 100 simulated samples of size 25, 50, or 100 from the model  $Y = 10 \sin(\pi X_1 X_2) + 30(X_3 - 0.5)^2 + 10X_4 + 5X_5 + \varepsilon$ , where  $X_i$ 's were distributed as  $U(0, 1)$  and  $\varepsilon$  was distributed as  $N(0, 1)$ . The numbers in parentheses indicate the proportion of significant CID / pCID values at  $\alpha = 0.05$  in 100 simulations.

	Average CID / pCID (sig. prop.)		
	$N = 25$	$N = 50$	$N = 100$
CID( $Y X_1$ )	0.0580 (0.34)	0.0542 (0.71)	0.0665 (0.98)
CID( $Y X_2$ )	0.0641 (0.50)	0.0667 (0.77)	0.0683 (0.98)
CID( $Y X_3$ )	0.0388 (0.16)	0.0353 (0.39)	0.0366 (0.83)
CID( $Y X_4$ )	0.1072 (0.82)	0.1034 (0.98)	0.1177 (1.00)
CID( $Y X_5$ )	0.0407 (0.25)	0.0365 (0.43)	0.0328 (0.69)
CID( $Y X_6$ )	0.0212 (0.06)	0.0145 (0.07)	0.0077 (0.02)
pCID( $Y X_1; X_4$ )	0.0608 (0.36)	0.0573 (0.75)	0.0644 (0.97)
pCID( $Y X_2; X_4$ )	0.0729 (0.55)	0.0704 (0.87)	0.0685 (0.99)
pCID( $Y X_3; X_4$ )	0.0427 (0.19)	0.0371 (0.44)	0.0322 (0.82)
pCID( $Y X_5; X_4$ )	0.0471 (0.23)	0.0359 (0.41)	0.0269 (0.65)
pCID( $Y X_6; X_4$ )	0.0270 (0.07)	0.0122 (0.04)	0.0015 (0.05)
pCID( $Y X_1; X_2, X_4$ )	0.0850 (0.45)	0.0820 (0.88)	0.0852 (0.99)
pCID( $Y X_3; X_2, X_4$ )	0.0624 (0.23)	0.0507 (0.41)	0.0398 (0.81)
pCID( $Y X_5; X_2, X_4$ )	0.0658 (0.19)	0.0503 (0.37)	0.0356 (0.66)
pCID( $Y X_6; X_2, X_4$ )	0.0463 (0.06)	0.0251 (0.04)	0.0088 (0.03)
pCID( $Y X_3; X_1, X_2, X_4$ )	0.0798 (0.23)	0.0719 (0.46)	0.0574 (0.84)
pCID( $Y X_5; X_1, X_2, X_4$ )	0.0789 (0.20)	0.0709 (0.48)	0.0531 (0.70)
pCID( $Y X_6; X_1, X_2, X_4$ )	0.0608 (0.08)	0.0451 (0.02)	0.0262 (0.03)
pCID( $Y X_5; X_1, X_2, X_3, X_4$ )	0.0753 (0.26)	0.0776 (0.50)	0.0721 (0.75)
pCID( $Y X_6; X_1, X_2, X_3, X_4$ )	0.0613 (0.04)	0.0565 (0.06)	0.0478 (0.05)
pCID( $Y X_6; X_1, X_2, X_3, X_4, X_5$ )	0.0464 (0.08)	0.0516 (0.07)	0.0513 (0.04)

### 3.2.2 The results of PCC and pPCC

When the Pearson's correlation coefficient (PCC) and the partial correlation coefficient (pPCC) were adopted to select relevant variables using the simulated data of Model (3.1), the explanatory variable  $X_4$  was the most linearly associated with the target variable  $Y$  by having the largest average value of  $PCC(Y, X_i)$  among all  $i = 1, \dots, 6$  regardless of the sample size (Table 3.4). About 88% to 100%  $PCC(Y, X_4)$  values were significantly not equal to zero according to their permutation p-values. The  $PCC(Y, X_1)$ ,  $PCC(Y, X_2)$ , and  $PCC(Y, X_5)$  values ranged from 0.2 to 0.4, values which were mostly identified as significant under the sample size  $N = 100$ . The proportion of significant  $PCC(Y, X_3)$  values in 100 simulations, however, was roughly consistent with the nominal significance level of  $\alpha = 0.05$ . To eliminate the impact from the dominant explanatory variable, the pPCC was proposed to quantify linear associations between a relatively minor explanatory variable to the target variable (Baba *et al.*, 2004). As illustrated in Table 3.4,  $X_2$ ,  $X_1$ , and  $X_5$  were sequentially selected according to the average values of pPCC, and  $pPCC(Y, X_5; X_4, X_2, X_1)$  were mostly significant.  $X_3$  was frequently discarded together with the irrelevant variable  $X_6$  in the variable selection process. This was expected due to the natural utilization of the PCC and the pPCC; they were specifically designed to detect linear association instead of association in general forms.



Table 3.4: Summary statistics of Pearson's correlation coefficients (PCC) and partial correlation coefficients (pPCC) based 100 simulated samples of size 25, 50, or 100 from the model  $Y = 10 \sin(\pi X_1 X_2) + 30(X_3 - 0.5)^2 + 10X_4 + 5X_5 + \varepsilon$ , where  $X_i$ 's were distributed as  $U(0, 1)$  and  $\varepsilon$  was distributed as  $N(0, 1)$ . The numbers in parentheses indicate the proportion of significant CID / pCID values at  $\alpha = 0.05$  in 100 simulations.

	Average PCC / pPCC (sig. prop.)		
	$N = 25$	$N = 50$	$N = 100$
PCC( $Y, X_1$ )	0.3401 (0.38)	0.3431 (0.73)	0.3759 (0.98)
PCC( $Y, X_2$ )	0.3845 (0.53)	0.3903 (0.77)	0.3815 (0.98)
PCC( $Y, X_3$ )	-0.0050 (0.02)	-0.0158 (0.07)	0.0020 (0.03)
PCC( $Y, X_4$ )	0.5691 (0.88)	0.5426 (0.98)	0.5586 (1.00)
PCC( $Y, X_5$ )	0.2645 (0.25)	0.2718 (0.45)	0.2753 (0.79)
PCC( $Y, X_6$ )	0.0008 (0.09)	0.0020 (0.06)	0.0073 (0.01)
pPCC( $Y, X_1; X_4$ )	0.4105 (0.54)	0.4221 (0.87)	0.4432 (1.00)
pPCC( $Y, X_2; X_4$ )	0.4629 (0.70)	0.4614 (0.94)	0.4548 (1.00)
pPCC( $Y, X_3; X_4$ )	-0.0063 (0.07)	0.0028 (0.07)	-0.0008 (0.05)
pPCC( $Y, X_5; X_4$ )	0.3158 (0.36)	0.3060 (0.56)	0.3204 (0.91)
pPCC( $Y, X_6; X_4$ )	-0.0006 (0.07)	-0.0042 (0.04)	0.0014 (0.03)
pPCC( $Y, X_1; X_2, X_4$ )	0.4759 (0.65)	0.4774 (0.93)	0.5063 (1.00)
pPCC( $Y, X_3; X_2, X_4$ )	-0.0207 (0.08)	0.0008 (0.08)	0.0034 (0.09)
pPCC( $Y, X_5; X_2, X_4$ )	0.3437 (0.36)	0.3499 (0.71)	0.3554 (0.96)
pPCC( $Y, X_6; X_2, X_4$ )	0.0066 (0.08)	-0.0006 (0.05)	-0.0091 (0.04)
pPCC( $Y, X_3; X_1, X_2, X_4$ )	-0.0208 (0.09)	0.0076 (0.10)	-0.0003 (0.11)
pPCC( $Y, X_5; X_1, X_2, X_4$ )	0.4189 (0.48)	0.4220 (0.86)	0.4140 (0.98)
pPCC( $Y, X_6; X_1, X_2, X_4$ )	0.0085 (0.05)	-0.0094 (0.07)	-0.0170 (0.01)
pPCC( $Y, X_3; X_1, X_2, X_4, X_5$ )	-0.0215 (0.11)	0.0060 (0.09)	0.0032 (0.13)
pPCC( $Y, X_6; X_1, X_2, X_4, X_5$ )	0.0088 (0.07)	0.0083 (0.03)	-0.0064 (0.01)
pPCC( $Y, X_6; X_1, X_2, X_3, X_4, X_5$ )	0.0054 (0.06)	0.0069 (0.09)	0.0031 (0.15)

### 3.3 *Arabidopsis* microarray data analysis

We exercised pCID to identify the genes that were associated with (or possibly regulate or be regulated by) a given transcription factor. The method was utilized to select gene signatures using *Arabidopsis Thaliana* (*Arabidopsis*) microarray dataset. The dataset contained the expression levels of *Arabidopsis* genes under cold stress, which can be downloaded from the *Arabidopsis* Information Resource (TAIR) database (Huala et al, 2001). This data originally consists of 22,810 probes and 52 samples (submission number ME00325) treated under cold stress (4 °C) after 0 (control), 0.5, 1, 3, 6, 12 or 24 hours (H). After normalized by the robust multichip average (RMA) method (Irizarry *et al.*, 2003) and log<sub>2</sub>-transformed with the BioConductor (Gentleman *et al.*, 2004) 'affyLmGUI' package (Wettenhall *et al.*, 2006), the expressions of all probes had to be tested by the analysis of variance (ANOVA). The probes having FDR < 0.001 under the time-course cold treatment were then further proceeded to CID/pCID analysis (Benjamini and Hochberg, 1995). Three C-repeat binding factors, *CBF1* (probe ID: 254074\_at), *CBF2* (probe ID: 254075\_at), and *CBF3* (probe ID:254066\_at), were all cold-responsive genes and were adopted as the explanatory variables *X*'s in CID/pCID demonstration while each of the other probes was treated as the target variable in our analysis.

The expression of C-repeat binding factor (CBF) genes in plants under different abiotic stresses has been extensively studied (Akhtar *et al.*, 2012). In *Arabidopsis*, three CBF genes (*CBF1*, *CBF2* and *CBF3*) were found to be active under cold stress (Gilmour *et al.*, 2004; Liu *et al.*, 1998). Here, the proposed CID/pCID methodology was exercised in studying cold-stress responsive regulation paths governed by three key regulatory proteins, CBF1, CBF2, and CBF3, at the transcriptional level using microarray gene expression data. There were 2,388 probes, including three probes of three CBF genes, identified as cold-responsive genes (ANOVA FDR < 0.001). Three CBF genes were further treated as the explanatory variable (*X*), and each one of the remaining 2,385 probes was treated as the target variable (*Y*) for CID/pCID analysis.

Among the 2,385 probes, 91% (2,177 probes) were significantly associated (CID/pCID *p*-values < 0.05) with at least one of the three CBF probes of interest in terms of their expression levels (Figure 3.3A). 26% (615 probes), 43% (939 probes), and 26% (623 probes) had the largest significant CID values given *CBF1*, *CBF2*, and *CBF3*, respectively. Only 431 probes had selected the second relevant CBF probes with

significant pCID values (pCID  $p$ -values  $< 0.05$ ); 192 out of 431 probes (45%) were related to both *CBF1* and *CBF2*, 79 (18%) were related to both *CBF2* and *CBF3*, and 160 (37%) were related to both *CBF1* and *CBF3* (Figure 3.3A). However, none of the 2,385 probes were associated with all three CBF probes by having all pCID with  $p$ -values  $\geq 0.05$  given any two CBF probes (Figure 3.3A).

The PCC/pPCC method identified fewer significant probes than the CID/pCID. Among the 2,385 probes, 78% (1,862 probes) were significantly associated (PCC/pPCC permutation  $p$ -values  $< 0.05$ ) with at least one of the three CBF probes of interest in terms of their expression levels (Figure 3.3B). However, 63% (1,169 probes) of the significant probes were found to be relevant to more than one of the three transcription factors; 105 probes were related to all three transcription factors. There were 1,849 probes commonly identified by both the CID/pCID and PCC/pPCC methods (Figure 3.3C). Five well known CBF target genes, *COR6.6* (246481\_s\_at), *COR78* (248337\_at), *COR47* (259570\_at), *COR15B* (263495\_at), and *COR15A* (263497\_at), were all commonly identified by both the CID/pCID and PCC/pPCC methods.

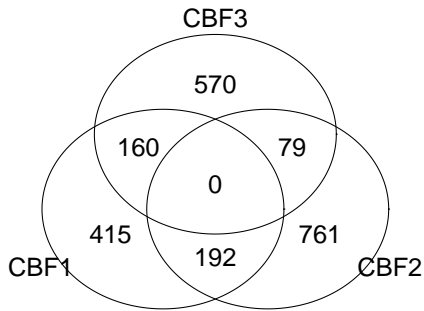
There were 328 and 13 probes, respectively, that were only identified by the CID/pCID or the PCC/pPCC method. This outcome implied, first, that PCC/pPCC was more sensitive (but maybe less specific) for identifying linear relationships than the CID/pCID method, and second, that the CID/pCID method identified nonlinear patterns of regulation of transcription factors to their target genes. More genes were identified as being significantly associated with more than two CBF TFs by the PCC/pPCC method, even though we initially expected the association to have been relatively weakened after removing the effect from the first identified CBF TF's.

Gene set enrichment analysis (Du *et al.*, 2010) was performed on 2,177, 1,862, and 1,849 probes identified as being associated with at least one of the three CBF probes by the CID/pCID method, the PCC/pPCC method, or by both, respectively. There were 154, 134, and 132 significant gene ontology (GO) accessions enriched (FDR  $< 0.01$ ), respectively, where 124 GO accessions were commonly identified (Figure 3.3D). Information for 29 GO accessions identified as being significantly enriched only by the CID/pCID method is listed in Table 3.5. We investigated further into two accessions: GO:0052544 (callose deposition in cell wall during defense response) and GO:0052482 (cell wall thickening during defense response); both accessions were identified through the same seven significant probes (264052\_at, 264873\_at, 262899\_at, 254270\_at, 253534\_at, 267392\_at, and 255378\_at), and all of

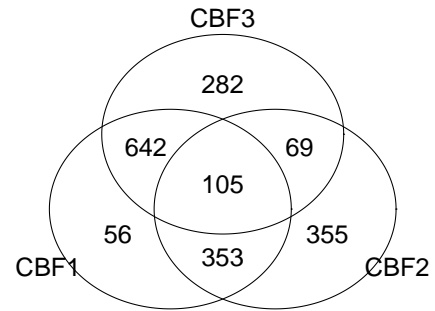




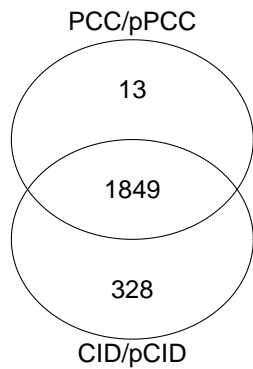
A. Cold responsive genes by CID/pCID



B. Cold responsive genes by PCC/pPCC



C. Cold responsive genes by CID/pCID and/or PCC/pPCC



D. Enriched GO accessions by CID/pCID and/or PCC/pPCC

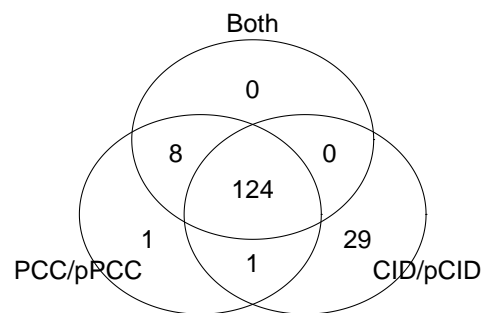


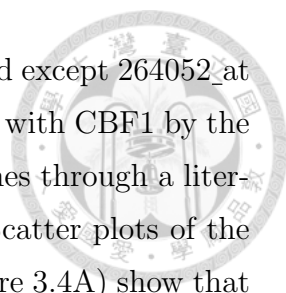
Figure 3.3: Venn diagrams of the 2,385 cold-responsive genes associated with three CBF transcription factors according to (A) the CID/pCID method, (B) the PCC/pPCC method, and (C) the CID/pCID method and/or the PCC/pPCC method. (D) Venn diagrams of the significantly enriched gene ontology accessions according to the CID/pCID method and/or the PCC/pPCC method.



Table 3.5: Information for 29 GO accessions identified as being significantly enriched according to CID/pCID significance.

Accession <sup>1</sup>	Type <sup>2</sup>	Description	FDR
GO:0016138*	P	glycoside biosynthetic process	0.0003
GO:0051179*	P	localization	0.0003
GO:0006810*	P	transport	0.0012
GO:0051234*	P	establishment of localization	0.0014
GO:0033036*	P	macromolecule localization	0.0024
GO:0052542	P	callose deposition during defense response	0.0031
GO:0007166	P	cell surface receptor linked signaling pathway	0.0033
GO:0033037	P	polysaccharide localization	0.0049
GO:0052545	P	callose localization	0.0049
GO:0044272*	P	sulfur compound biosynthetic process	0.0050
GO:0007275*	P	multicellular organismal development	0.0070
GO:0007167	P	enzyme linked receptor protein signaling pathway	0.0073
GO:0007169	P	transmembrane receptor protein tyrosine kinase signaling pathway	0.0073
GO:0010200*	P	response to chitin	0.0075
GO:0052544	P	callose deposition in cell wall during defense response	0.0084
GO:0052482	P	cell wall thickening during defense response	0.0084
GO:0010876*	P	lipid localization	0.0095
GO:0032555*	F	purine ribonucleotide binding	0.0014
GO:0032553*	F	ribonucleotide binding	0.0014
GO:0000166*	F	nucleotide binding	0.0019
GO:0032559*	F	adenyl ribonucleotide binding	0.0027
GO:0017076*	F	purine nucleotide binding	0.0032
GO:0005524*	F	ATP binding	0.0042
GO:0004713	F	protein tyrosine kinase activity	0.0057
GO:0010011	F	auxin binding	0.0061
GO:0005506	F	iron ion binding	0.0062
GO:0001882*	F	nucleoside binding	0.0071
GO:0001883*	F	purine nucleoside binding	0.0071
GO:0030554*	F	adenyl nucleotide binding	0.0071

<sup>1</sup>Eighteen accessions containing the 42 genes associated with more than one CBF TFs according to CID/pCID are marked “\*”. <sup>2</sup>Accession types: biological process (P), cellular component (C), and molecular function (F).



them were also identified as significant by the PCC/pPCC method except 264052\_at (AT2G22330) and 253534\_at (AT4G31500); both were associated with *CBF1* by the CID/pCID method and were confirmed to be cold-responsive genes through a literature search (Fowler and Thomashow, 2002; Lee *et al.*, 2005). Scatter plots of the expressions of these two probes to the expressions of *CBF1* (Figure 3.4A) show that only moderate linear relationships exist when the log<sub>2</sub> expression levels of *CBF1* were greater than 7; the scattered patterns when *CBF1* lowly express weakened the linearity ( $r = -0.13$  and  $-0.14$ , respectively). By plotting the average log<sub>2</sub> expression levels (Figure 3.4B), we observed that the expressions of 264052\_at and 253534\_at descended along with those of *CBF1* from 3H to 24H after cold treatment.

Conceptually, the CID values are computed from the cumulative discrepancies between the marginal and conditional distributions. By comparing such discrepancies observed from each sample, we are able to check in which sample subsets a stronger association between the predictor and the target variables can be observed. Figure 3.4C shows the percentages of the sample subsets that contributed to the association of the CBF TFs with the significant genes. The dashed horizontal line represents the value  $1/26 = 0.038$  when all 26 tissues  $\times$  times  $\times$  treatments combinations equally contributed to the CID value. The information provided by the expression of *CBF1* from shoot tissue at 24H after treatment, for example, contributed more than 15% of the significant  $\text{CID}(264052\_at|CBF1)$  and  $\text{CID}(253534\_at|CBF1)$ , respectively. More specifically, 264052\_at and 253534\_at mostly had relatively large expression values when the expression levels of *CBF1* were around the range observed from shoot tissue at 24 hour after treatment (from Figure 3.4A, or from Figure 3.4D showing the conditional CDF's due to samples under 24H cold treatment [yellow dashed lines] are above the marginal CDF [black solid line]). The information provided for 264052\_at by *CBF1* from root and shoot tissues at one hour after treatment also largely contributed the CID value, but 264052\_at had relatively high expression levels in shoot tissues but relatively low expression levels in root tissues. This implies that the contributions of the sub-samples to the CID values are capable of indicating the sample-specific gene-gene interactions.

Furthermore, 42 genes were associated with more than one CBF TF according to the CID/pCID method but were not identified as significant by the PCC/pPCC method. These genes were contributed to eighteen GO accessions enriched only by the CID/pCID method (Table 3.5), where 253114\_at (AT4G35860) associated with both *CBF1* and *CBF3* contributed to the enrichment of 8 GO accessions. The

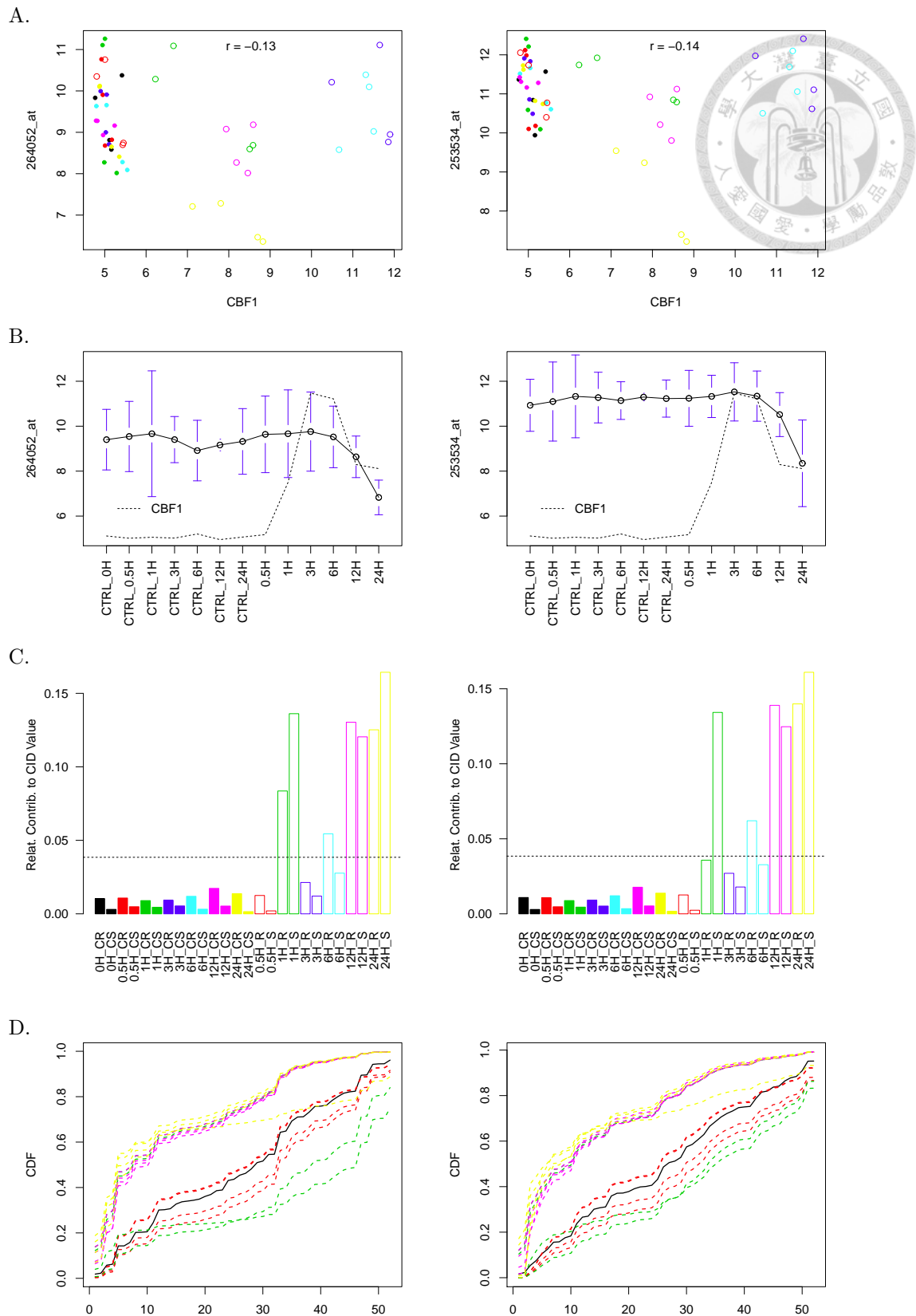
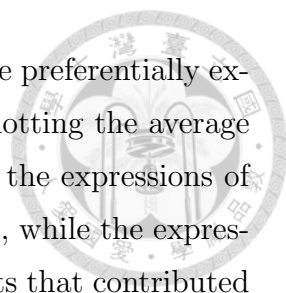


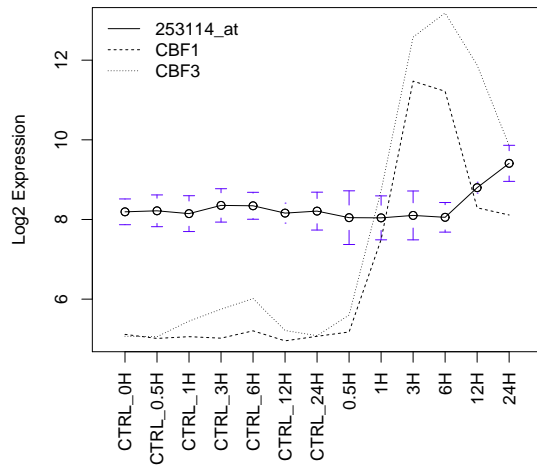
Figure 3.4: Expression profiles and CID/pCID inferences of 264052\_at and 253534\_at based on expression levels of *CBF1*. (A) Scatter plots of log<sub>2</sub> expression levels. (B) Averages and standard deviations of log<sub>2</sub> expression levels over time under control (CTRL) or cold treatments. (C) Contribution to CID value by different sub-samples. C: control; S: shoot; R: root. The dashed horizontal line indicates the nominal value 1/26. (D) Marginal CDF (black solid line) and conditional CDF's under 0.5H\_R, 0.5H\_S (red dashed lines), 1H\_R, 1H\_S (green dashed lines), 12H\_R, 12H\_S (pink dashed lines), 24H\_R, and 24H\_S (yellow dashed lines).



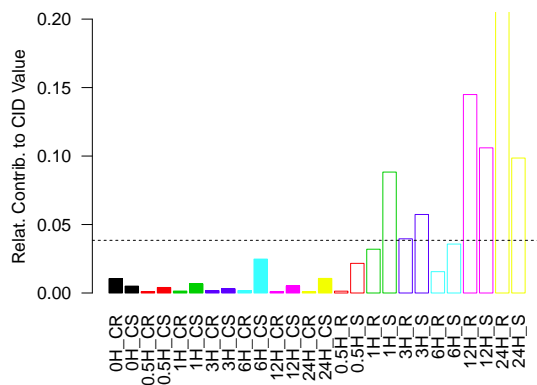
gene corresponding to 253114\_at was previously reported as a gene preferentially expressed in cold stored peach fruits (Tittarelli *et al.*, 2009). By plotting the average log2 expression levels over time (Figure 3.5A), we observed that the expressions of *CBF1* and *CBF3* decreased from 6H to 24H after cold treatment, while the expression of 253114\_at increased. The percentages of the sample subsets that contributed to the association of 253534\_at with *CBF1* and *CBF3* are shown in Figure 3.5B and Figure 3.5C. The information provided by the expression of *CBF1* at 24H after treatment contributed most to the significance of  $CID(253114\_at|CBF1)$ , and the information provided by the expression of *CBF3* at 3H after treatment contributed most to the significance of  $pCID(253114\_at|CBF3;CBF1)$ . A minor negative correlation between *CBF3* and 253114\_at was also observed in the control samples from 6H to 24H. This feature was captured by the discrepancy between the marginal and conditional distributions at 6H after treatment in the control shoot sample when calculating  $pCID(253114\_at|CBF3;CBF1)$  (Figure 3.5C). Further experiments can be conducted to confirm these hypotheses.



A.



B.



C.

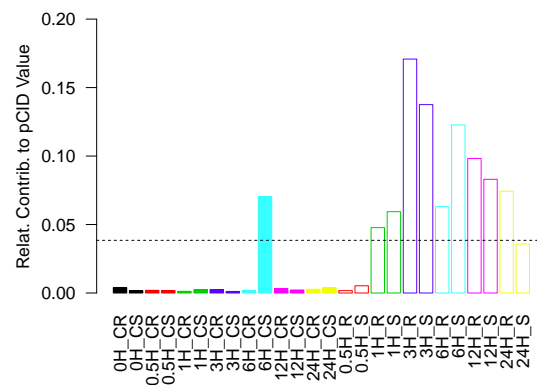


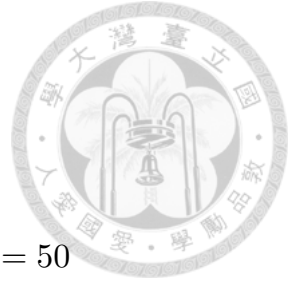
Figure 3.5: Expression profiles and CID/pCID inferences of 253114\_at based on expression levels of *CBF1* and *CBF3*. (A) Averages and standard deviations of log<sub>2</sub> expression levels over time under control (CTRL) or cold treatments. (B) Contribution to CID(253114\_at|*CBF1*) and (C) pCID(253114\_at|*CBF3*; *CBF1*) by different sub-samples. C: control; S: shoot; R: root. The dashed horizontal line indicates the nominal value 1/26.

### 3.4 Discussion

The CID values of  $Y$  given either one or two predictors provided hints regarding how to guess about the approximate pCID values. For example,  $\text{pCID}(Y|X_1; X_4)$  is approximately  $(0.1747 - 0.1176)/(1 - 0.1176) = 0.065$  and  $\text{pCID}(Y|X_6; X_4)$  is approximately  $(0.1191 - 0.1176)/(1 - 0.1176) = 0.002$  (see Table 3.1). The latter is much smaller than the former, reflecting their differing magnitudes of dependency. After eliminating the impact from the more dominant variables, the signals from the minor variables were enlarged and the pCID values were gradually increased as the number of conditioning variables was increased.

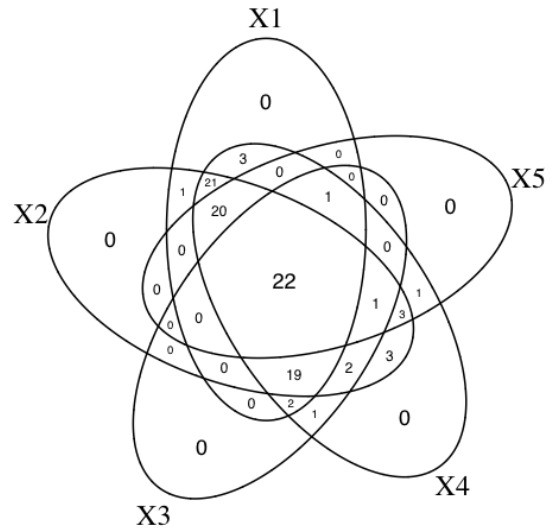
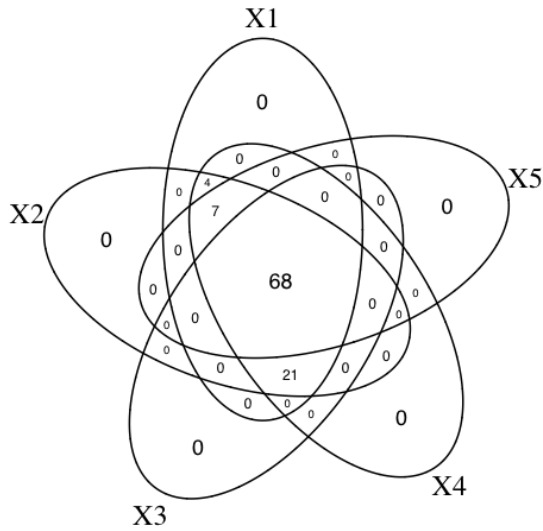
The order of the variables declared relevant also provided hints about the order of priority for statistical dependence. Linearity was superior to nonlinearity because  $X_4$  was favored over  $X_1$  and  $X_2$  even though  $10X_4$  and  $10\sin(\pi X_1 X_2)$  contributed the same range of  $Y$  in Model (3.1). But the influence of  $X_2$  (or  $X_1$ ) was stronger than that of  $X_5$ , which had only half the impact of  $X_4$  on  $Y$  in the model.  $X_3$  and  $X_5$  having similar CID and pCID values (see Table 3.1) but the range of  $30(X_3 - 0.5)^2$  and  $5X_5$  being  $[0, 7.5]$  and  $[0, 5]$ , respectively, means that  $X_5$  was 1.5 times ‘more influential’ on  $Y$  than  $X_3$ . Therefore, pCID values can serve as indicators for or can even quantify different types of curvilinearity in regard to statistical dependence.

With a relatively large sample size ( $N = 100$ ), 96% of the simulations correctly selected more than four of five relevant variables, while the irrelevant variable  $X_6$  was falsely included in only three simulations (Figure 3.6A). Otherwise, 22% of the simulations under the moderate sample size (e.g.,  $N = 50$ ) picked all five relevant variables; 41% of the simulations picked four relevant variables, where  $X_4$  was never missed but  $X_3$  and  $X_5$  were missed in about 20% of the simulations (Figure 3.6B). Also about 20% of the simulations claimed significance only for  $X_1$ ,  $X_2$ , and  $X_4$  (Figure 3.6B). For a small sample size ( $N = 25$ ), CID / pCID lost sensitivities in finding  $X_5$  (79% missed),  $X_3$  (78% missed),  $X_1$  (51% missed),  $X_2$  (44% missed), and  $X_4$  (17% missed) (Figure 3.6C). But  $X_6$  was selected in 8% of the simulations, which is about the nominal  $\alpha = 0.05$ .



A.  $N = 100$

B.  $N = 50$



C.  $N = 25$

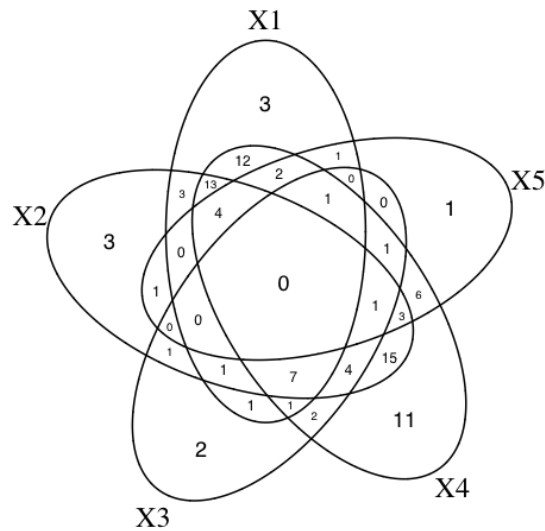


Figure 3.6: Number of the relevant variable  $X_i$  ( $i = 1, 2, 3, 5, 6$ ) being selected in 100 simulated samples of size (A) 100, (B) 50, or (C) 25 from the model  $Y = 10 \sin(\pi X_1 X_2) + 30(X_3 - 0.5)^2 + 10X_4 + 5X_5 + \epsilon$ , where  $X_i$ 's were distributed as  $U(0, 1)$  and  $\epsilon$  was distributed as  $N(0, 1)$ .





## Chapter 4

# Application to gene regulatory network

The gene regulation events under certain condition serve as small blocks to the entire gene regulation network (GRN), which may be reconstructed by connecting multiple regulation modules. An inferred GRN can therefore provide insights into the relationships between genes of interest by experiments and the understanding of biological functions with complex biological phenomena. More specifically, an inferred GRN consisting of the nodes (representing genes) and the edges (representing significant gene-gene interaction) reflects the gene regulation events that may concurrently or sequentially occur under the condition of study. In this study, we focus on the inference of GRN using the results of microarray experiments. It is usually achieved by (1) identifying a pair of significantly associated genes, (2) elongating the regulation path from the gene pair, and then (3) assembling all identified paths to form the complex GRN (Figure 4.1).

This study aims to infer the causality in a GRN using CID. A causal connection between a pair of nodes means one is the origin (source) and the other is the consequence (target) in the association. Such cause and effect relationship is usually expected when studying the relationship between a transcription factor (TF) and its target genes and is usually indicated as a directed edge in the network. Compared to co-expression GRN (i.e., network with undirected edges), a cause-and-effect GRN requires more information to put the direction on the edge. The direction is typically assigned according to known biological evidences, which may not be available at all time. In this study, we utilize the asymmetric property of CID (i.e.,  $CID(Y|X)$  is not necessarily equal to  $CID(X|Y)$ ) to distinguish not only the associated gene pairs but the causes / effects in a gene regulation event. Asymmetry is a very unique feature of CID whereas the some conventional methods, including PCC, pPCC and MI,

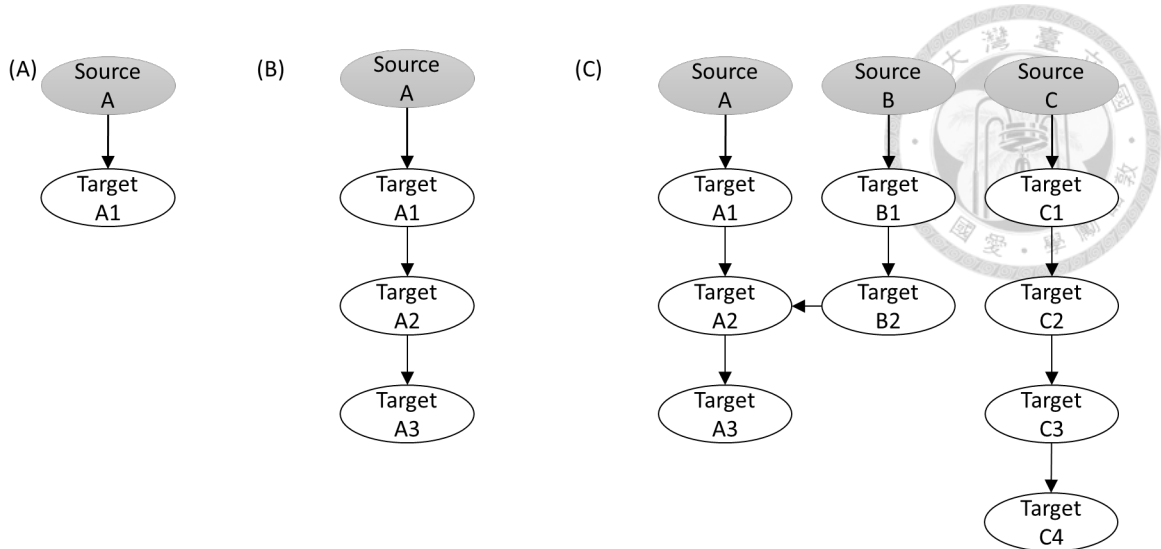


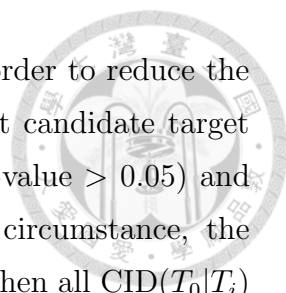
Figure 4.1: Diagram of gene regulatory network inference workflow. (A) Identification of a significantly associated gene pair. (B) Regulation path elongation. (C) Assembly of all identified regulation paths.

provide symmetric results when considering the association between two variables. More specifically, the gene  $Y$  is designated as the source and gene  $X$ , the target, in the GRN if  $\text{CID}(Y|X) > \text{CID}(X|Y)$ .

The pCID method could identify relevant genes in the elongation step. Ideally, a proper stepwise procedure iteratively picks the relevant genes according to its magnitude of association to the target until no more gene would significantly increase the amount of association. For example, in Figure 4.1B,  $\text{CID}(\text{Source A}|\text{Target A1})$  would be significant while we also expect a significant  $\text{CID}(\text{Source A}|\text{Target A1, Target A2})$  but a insignificant  $\text{CID}(\text{Source A}|\text{Target A1, } X)$  given an irrelevant gene  $X$ . However, due to the dominant effect of the most influential gene, i.e., Target A1, in the first step,  $\text{CID}(\text{Source A}|\text{Target A1, } X)$  were mostly significant (see Section 3). The pCID resolves this problem by decomposing only the information of the target variable which was not explained by the first predictor.

## 4.1 Construction of gene regulatory network by CID/pCID

The inference of GRN has three steps (Figure 4.1). However, due to the dramatic amount of genes simultaneously monitored in a microarray experiment, we develop the following heuristic approach for the first two steps which were illustrated with Figure 4.2. Given a source gene  $T_0$ ,  $\text{CID}(T_0|T_i)$  for one of the candidate target genes,  $T_i$ , was computed in the first step. The candidate target genes may be all



other genes in the same microarray dataset or user-defined. In order to reduce the computation of the programming, we eliminated some irrelevant candidate target genes which caused the  $CID(T_0|T_i)$  values to be insignificant ( $p$ -value  $> 0.05$ ) and which were not proceeded to the following steps. Under the circumstance, the source gene  $T_0$  was discarded as the origin of a regulation path when all  $CID(T_0|T_i)$  values were insignificant in the first run. Otherwise, if  $CID(T_0|T_{(1)})$  had the single smallest significant  $p$ -value among the results from all candidate target genes, we connected the source gene  $T_0$  and the target gene  $T_{(1)}$ . Provided that there were more than one  $CID(T_0|T_i)$  value had the smallest significant  $p$ -value, we selected  $T_{(1)}$  which had the maximum of these  $CID(T_0|T_i)$  value. The decision-making about the direction between the source gene  $T_0$  and the target gene  $T_{(1)}$  was based on comparing the significance between  $CID(T_0|T_{(1)})$  and  $CID(T_{(1)}|T_0)$ . If  $CID(T_0|T_{(1)})$  was more significant than  $CID(T_{(1)}|T_0)$  or if these two  $CID$  values had equal  $p$ -value and the  $CID(T_0|T_{(1)})$  value was larger than  $CID(T_{(1)}|T_0)$  value, the direction was from  $T_0$  to  $T_{(1)}$ ; otherwise, the direction was from  $T_{(1)}$  to  $T_0$ . The gene pair  $(T_0, T_{(1)})$  was proceeded to the elongation step.

In the elongation step,  $pCID(T_0|T_j; T_{(1)})$  and  $pCID(T_{(1)}|T_j; T_0)$  were computed for one of the remaining candidate target genes,  $T_j$ , to identify the second relevant target gene,  $T_{(2)}$  (Figure 4.2). Suppose that all  $pCID(T_0|T_j; T_{(1)})$  and  $pCID(T_{(1)}|T_j; T_0)$  values were insignificant, the regulation path would stop and the network was with two nodes  $(T_0, T_{(1)})$ . In other respects, the process was continued and there were two routes to connect the regulation path. Provided that there were more than one  $pCID(T_0|T_j; T_{(1)})$  or  $pCID(T_{(1)}|T_j; T_0)$  value had the smallest significant  $p$ -value among the results of the  $pCID(T_0|T_j; T_{(1)})$  and  $pCID(T_{(1)}|T_j; T_0)$  from all remaining candidate target genes, we selected  $T_{(2)}$  which had the maximum of these  $pCID(T_0|T_j; T_{(1)})$  and  $pCID(T_{(1)}|T_j; T_0)$  values. One of these routes was that we connected the gene  $T_0$  and  $T_{(2)}$ , if  $T_{(2)}$  was selected as a result of the  $pCID(T_0|T_{(2)}; T_{(1)})$  value. The decision of the direction by  $pCID$  values was similar to the previous resolution by  $CID$  values. The direction was from  $T_0$  to  $T_{(2)}$ , if  $pCID(T_0|T_{(2)}; T_{(1)})$  was more significant than  $pCID(T_{(2)}|T_0; T_{(1)})$  or if these two  $pCID$  values had equal  $p$ -value and the  $pCID(T_0|T_{(2)}; T_{(1)})$  value was larger than  $pCID(T_{(2)}|T_0; T_{(1)})$  value; or from  $T_{(2)}$  to  $T_0$ , otherwise. The other route was that we connected the gene  $T_{(1)}$  and  $T_{(2)}$ , if  $T_{(2)}$  was selected as a result of the  $pCID(T_{(1)}|T_{(2)}; T_0)$  value. The direction was from  $T_{(1)}$  to  $T_{(2)}$ , if  $pCID(T_{(1)}|T_{(2)}; T_0)$  was more significant than  $pCID(T_{(2)}|T_{(1)}; T_0)$  or if these two  $pCID$  values had equal  $p$ -value and the  $pCID(T_{(1)}|T_{(2)}; T_0)$  value was

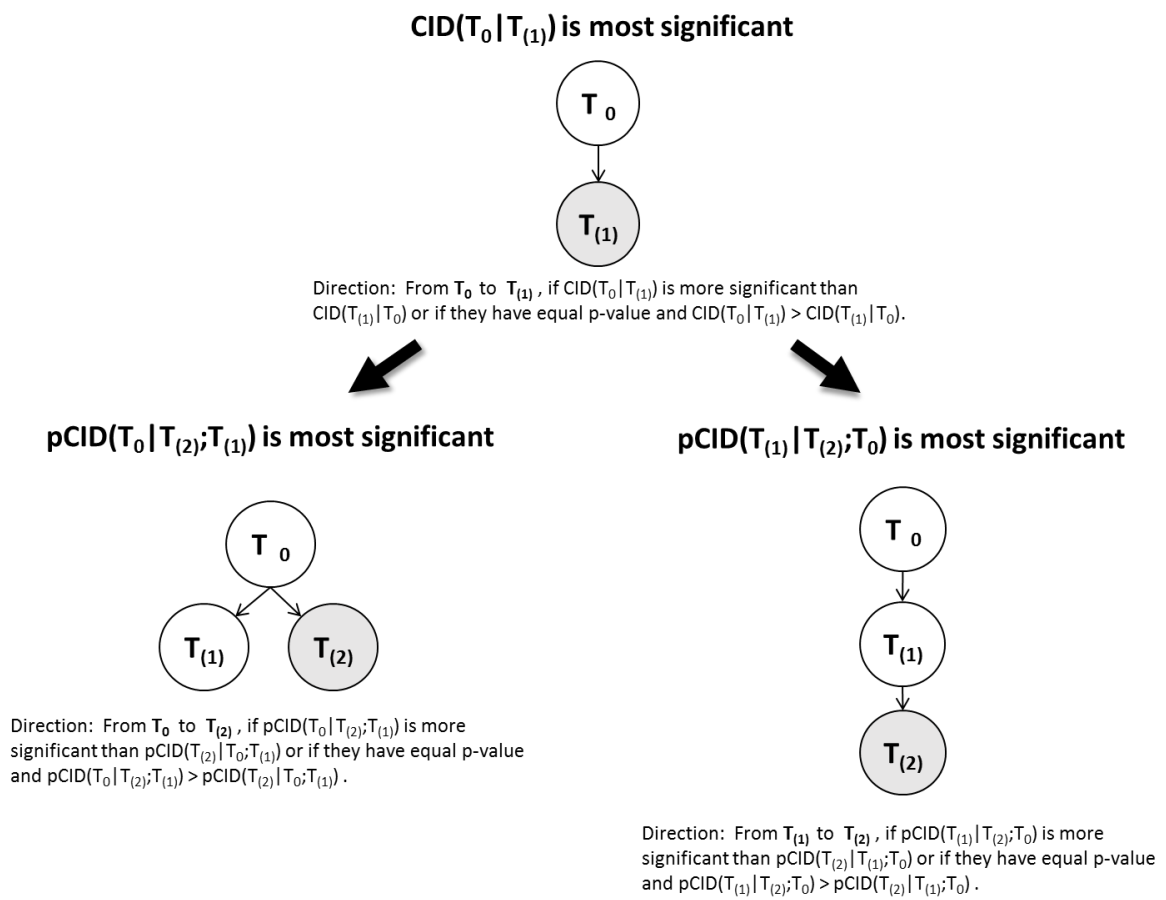


Figure 4.2: Illustration of the heuristic approach for regulation path elongation.

larger than  $\text{pCID}(T_{(2)}|T_{(1)}; T_0)$  value; or from  $T_{(2)}$  to  $T_{(1)}$ , otherwise. This finished the first run of the elongation.

Furthermore, we explain the next steps of GRN construction. In the  $r$ th run ( $r \geq 2$ ) of the elongation, all possible values of  $\text{pCID}(S|T_j; \{T_0, T_{(1)}, \dots, T_{(r)}\} \setminus S)$  for one of the remaining candidate genes,  $T_j$ , and  $S \in \{T_0, T_{(1)}, \dots, T_{(r)}\}$  were computed. Suppose that all  $\text{pCID}(S|T_j; \{T_0, T_{(1)}, \dots, T_{(r)}\} \setminus S)$  values were insignificant, the regulation path would stop and the network was with  $r + 1$  nodes ( $T_0, T_{(1)}, \dots, T_{(r)}$ ). Provided that there were more than one  $\text{pCID}(S|T_j; \{T_0, T_{(1)}, \dots, T_{(r)}\} \setminus S)$  value had the smallest significant  $p$ -value among the results of the  $\text{pCID}(S|T_j; \{T_0, T_{(1)}, \dots, T_{(r)}\} \setminus S)$  from all remaining candidate target genes, we selected  $T_{(r+1)}$  which had the maximum of these  $\text{pCID}(S|T_j; \{T_0, T_{(1)}, \dots, T_{(r)}\} \setminus S)$  value and connected the target gene  $S$  and  $T_{(r+1)}$ . The direction was from  $S$  to  $T_{(r+1)}$ , if  $\text{pCID}(S|T_{(r+1)}; \{T_0, T_{(1)}, \dots, T_{(r)}\} \setminus S)$  was more significant than  $\text{pCID}(T_{(r+1)}|S; \{T_0, T_{(1)}, \dots, T_{(r)}\} \setminus S)$  or if these two  $\text{pCID}$  values had equal  $p$ -value and the  $\text{pCID}(S|T_{(r+1)}; \{T_0, T_{(1)}, \dots, T_{(r)}\} \setminus S)$  value was larger than the  $\text{pCID}(T_{(r+1)}|S; \{T_0, T_{(1)}, \dots, T_{(r)}\} \setminus S)$  value; or from  $T_{(r+1)}$  to  $S$ , otherwise. The whole elongation process was continued until all of the  $\text{pCID}(S|T_j; \{T_0, T_{(1)}, \dots, T_{(e)}\} \setminus S)$  values in the  $e$ th run of the elongation were insignificant ( $p$ -value  $> 0.05$ ). The resulting network would contain  $e + 1$  nodes ( $T_0, T_{(1)}, \dots, T_{(e)}$ ). For example, Figure 4.3 illustrates one of the GRN construction results. Let  $T_0$  be the source gene and the other genes be the target genes. First (Step (0) in Figure 4.3), we computed all CID values of  $T_0$  given one of the target genes, and then  $\text{CID}(T_0|T_{(1)})$  had the most significant  $p$ -value, we connected the  $T_0$  and  $T_{(1)}$  with the direction was from  $T_0$  to  $T_{(1)}$  when the value of  $\text{CID}(T_0|T_{(1)}) > \text{CID}(T_{(1)}|T_0)$ . Second (Step (1)), we selected the target gene,  $T_{(2)}$ , which might be connected with  $T_0$  or  $T_{(1)}$ . Therefore, we computed the  $\text{pCID}(T_0|T_j; T_{(1)})$  and  $\text{pCID}(T_{(1)}|T_j; T_0)$ , where  $T_j$  was one of the remaining genes. The result was that  $\text{pCID}(T_0|T_{(2)}; T_{(1)})$  had the most significant  $p$ -value and  $T_{(2)}$  was connected with  $T_0$  from  $T_0$  to  $T_{(2)}$  when  $\text{pCID}(T_0|T_{(2)}; T_{(1)}) > \text{pCID}(T_{(2)}|T_0; T_{(1)})$  value. In Step (2), the next selected gene,  $T_{(3)}$ , could be connected with  $T_0$  or  $T_{(1)}$  or  $T_{(2)}$ . We computed the  $\text{pCID}(T_0|T_j; T_{(1)}, T_{(2)})$ ,  $\text{pCID}(T_{(1)}|T_j; T_0, T_{(2)})$  and  $\text{pCID}(T_{(2)}|T_j; T_0, T_{(1)})$ , where  $T_j$  was one of the remaining genes. The result was that  $\text{pCID}(T_0|T_{(3)}; T_{(1)}, T_{(2)})$  had the most significant  $p$ -value and  $T_{(3)}$  was connected with  $T_0$  from  $T_{(3)}$  to  $T_0$  when  $\text{pCID}(T_{(3)}|T_0; T_{(1)}, T_{(2)}) > \text{pCID}(T_0|T_{(3)}; T_{(1)}, T_{(2)})$ . In Step (3), the chosen target gene,  $T_{(4)}$ , would be connected with one of the prior selected genes ( $T_0, T_{(1)}, T_{(2)}$  and  $T_{(3)}$ ). We computed the  $\text{pCID}(T_0|T_j; T_{(1)}, T_{(2)}, T_{(3)})$ ,  $\text{pCID}(T_{(1)}|T_j; T_0, T_{(2)}, T_{(3)})$ ,

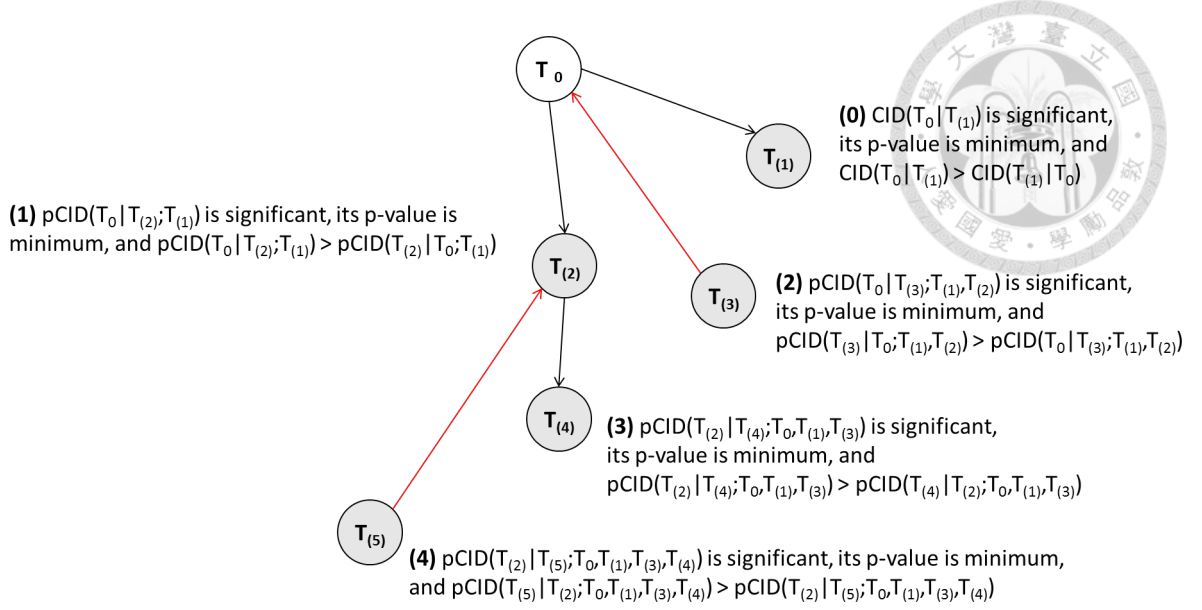
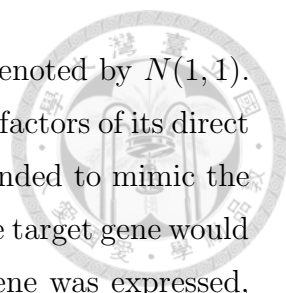


Figure 4.3: Illustration of the simple example for regulation path elongation used by CID/pCID method.

$\text{pCID}(T_{(2)}|T_j; T_0, T_{(1)}, T_{(3)})$  and  $\text{pCID}(T_{(3)}|T_j; T_0, T_{(1)}, T_{(2)})$ , where  $T_j$  was one of the remaining genes. Therefore the  $\text{pCID}(T_{(2)}|T_{(4)}; T_0, T_{(1)}, T_{(3)})$  had the most significant  $p$ -value and  $T_{(4)}$  was connected with  $T_{(2)}$  from  $T_{(2)}$  to  $T_{(4)}$  when  $\text{pCID}(T_{(2)}|T_{(4)}; T_0, T_{(1)}, T_{(3)}) > \text{pCID}(T_{(4)}|T_{(2)}; T_0, T_{(1)}, T_{(3)})$ . In Step (4), the chosen target gene,  $T_{(5)}$ , would be connected with one of the previous selected genes ( $T_0, T_{(1)}, T_{(2)}, T_{(3)}$  and  $T_{(4)}$ ). We computed the  $\text{pCID}(T_0|T_j; T_{(1)}, T_{(2)}, T_{(3)}, T_{(4)})$ ,  $\text{pCID}(T_{(1)}|T_j; T_0, T_{(2)}, T_{(3)}, T_{(4)})$ ,  $\text{pCID}(T_{(2)}|T_j; T_0, T_{(1)}, T_{(3)}, T_{(4)})$ ,  $\text{pCID}(T_{(3)}|T_j; T_0, T_{(1)}, T_{(2)}, T_{(4)})$  and  $\text{pCID}(T_{(4)}|T_j; T_0, T_{(1)}, T_{(2)}, T_{(3)})$ , where  $T_j$  was one of the remaining genes. Therefore the  $\text{pCID}(T_{(2)}|T_{(5)}; T_0, T_{(1)}, T_{(3)}, T_{(4)})$  had the most significant  $p$ -value and  $T_{(5)}$  was connected with  $T_{(2)}$  from  $T_{(5)}$  to  $T_{(2)}$  when  $\text{pCID}(T_{(5)}|T_{(2)}; T_0, T_{(1)}, T_{(3)}, T_{(4)}) > \text{pCID}(T_{(2)}|T_{(5)}; T_0, T_{(1)}, T_{(3)}, T_{(4)})$ . In the next step, we wanted to find the next linked gene  $T_{(6)}$  but all of  $\text{pCID}(S|T_j; \{T_0, T_{(1)}, \dots, T_{(5)}\} \setminus S)$  values were insignificant ( $p$ -value  $> 0.05$ ), where  $S$  was one of these previous selected genes,  $T_0, T_{(1)}, T_{(2)}, T_{(3)}, T_{(4)}$  and  $T_{(5)}$ .

## 4.2 Simulation study

The proposed procedure of GRN inference was examined in the simulation study. A pseudo network with six nodes (genes) was generated according to normal mixture model (Figure 4.4). It contained one source node (A11), four target nodes (A21, A22, A31 and A32), and one node (B) independent to the others. The expression levels of nodes A11 and B were randomly generated from the Normal distribution



with mean and standard deviation both equal to 1, which was denoted by  $N(1, 1)$ . The expression levels of the target nodes would be affected by two factors of its direct source: the expression level and the binding efficiency. This intended to mimic the occasions (1) the transcription factor was not expressed so that the target gene would not be regulated by the source gene, and (2) even the source gene was expressed, the target gene may still not be regulated by the source gene due to various binding efficiency of the transcription factor. Let  $S$  and  $T$  denote the direct source and the target gene, respectively. In the simulated network (Figure 4.4), A11 was the direct source of  $\{A21, A22\}$  and A21 was the direct source of  $\{A31, A32\}$ . If the binding efficiency for this pair of  $S$  and  $T$  was set to be  $100b\%$ , then  $100(1-b)\%$  of the objects in the sample were not affected by the expression level of  $S$  and their expression levels were generated from  $N(-1, 0.25)$ . The binding efficiency ( $b$ ) for  $\{A11, A21\}$ ,  $\{A11, A22\}$ ,  $\{A21, A31\}$ , and  $\{A21, A32\}$  were 0.9, 0.7, 0.9, and 0.8, respectively. For the  $100b\%$  objects that the regulation did take place, if the expression level of  $S$  in the  $i$ th sample was  $s_i$ , the expression level of the  $i$ th sample was randomly generated from  $N(s_i, 0.25)$  if  $s_i > 0$  and from  $N(-1, 0.25)$  if  $s_i < 0$  (meaning  $S$  was not expressed). Based on statistical theory, the approximate proportions of gene expressions of the target gene actually determined by the expression levels of the source gene were indicated next to the arrows in Figure 4.4. The inference process of the proportions of gene expressions of the target gene was showed in Appendix A. The pseudo network was replicated 100 times with sample size  $N = 25, 50$  and 100.

Simulation setup

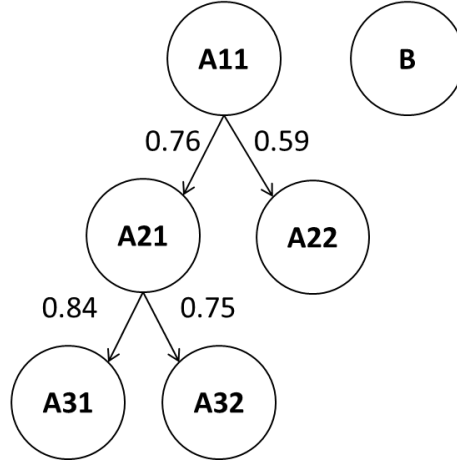


Figure 4.4: Pseudo network for the simulation study. The numbers next to the arrows illustrate the proportions of the objects in the sample that the expressions of the target node actually determined by the expressions of the source node.

A pseudo network with six nodes (genes) was generated to assess the proposed procedure of GRN inference (Figure 4.4). Two source genes, A11 and B, were predetermined. The CID and pCID values as well as their  $p$ -values for a particular simulation under sample size  $N = 50$  are shown in Table 4.1 for demonstration of network reconstruction. Starting from A11, the  $\text{CID}(\text{A11}|\text{B})$  value was insignificant ( $p$ -value:  $0.4136 > 0.05$ ), hence the node B did not exist in the following steps. Then the results showed  $\text{CID}(\text{A11}|\text{A21})$ ,  $\text{CID}(\text{A11}|\text{A22})$ ,  $\text{CID}(\text{A11}|\text{A31})$  and  $\text{CID}(\text{A11}|\text{A32})$  had the minimum  $p$ -value (0.0010) and  $\text{CID}(\text{A11}|\text{A22})$  value (0.2028) was the maximum of these CID values, so that A22 would be selected as the first node connected to A11. Because  $\text{CID}(\text{A11}|\text{A22})$  and  $\text{CID}(\text{A22}|\text{A11})$  had the same significant  $p$ -value (0.0010) and  $\text{CID}(\text{A11}|\text{A22})$  value (0.2028) was larger than  $\text{CID}(\text{A22}|\text{A11})$  value (0.1791), the direction was set from A11 to A22. The computation of  $\text{pCID}(\text{A11}|x; \text{A22})$  and  $\text{pCID}(\text{A22}|x; \text{A11})$  for another gene  $x$  followed and resulted in the selection of A21 as the second node connected to A11 due to that  $\text{pCID}(\text{A11}|\text{A21}; \text{A22})$  had the smallest  $p$ -value (0.0010) and the largest pCID value (0.1013). The direction was set from A11 to A21 because  $\text{pCID}(\text{A11}|\text{A21}; \text{A22})$  had the same significant  $p$ -value (0.0010) as  $\text{pCID}(\text{A21}|\text{A11}; \text{A22})$  and its value (0.1013) was larger than  $\text{pCID}(\text{A21}|\text{A11}; \text{A22})$  value (0.0934). Similarly, the third and fourth target, A31 and A32, was selected based on  $\text{pCID}(\text{A21}|\text{A31}; \text{A11}, \text{A22})$  and  $\text{pCID}(\text{A21}|\text{A32}; \text{A11}, \text{A22}, \text{A31})$ ; both A31 and A32 was connected from A21 due to  $\text{pCID}(\text{A21}|\text{A31}; \text{A11}, \text{A22})$  was equal significant ( $p$ -value: 0.0010) to and has larger value than





Table 4.1: The estimated CID and pCID values in one of the 100 simulations with sample size  $N = 50$ .

CID/pCID	Estimate ( $p$ -value)	CID/pCID	Estimate ( $p$ -value)
CID(A11 A21)	0.1936 (0.0010)		
CID(A11 A22)	<b>0.2028 (0.0010)</b>	CID(A22 A11)	0.1791 (0.0010)
CID(A11 A31)	0.1612 (0.0010)		
CID(A11 A32)	0.1281 (0.0010)		
CID(A11 B)	0.0129 (0.4136)		
pCID(A11 A21;A22)	<b>0.1013 (0.0010)</b>	PCID(A21 A11;A22)	0.0934 (0.0010)
pCID(A11 A31;A22)	0.0639 (0.0020)		
pCID(A11 A32;A22)	0.0534 (0.0010)		
pCID(A22 A21;A11)	0.0582 (0.0060)		
pCID(A22 A31;A11)	0.0446 (0.0100)		
pCID(A22 A32;A11)	0.0500 (0.0090)		
pCID(A11 A31;A21,A22)	0.0097 (0.2208)		
pCID(A11 A32;A21,A22)	0.0130 (0.1858)		
pCID(A21 A31;A11,A22)	<b>0.1131 (0.0010)</b>	pCID(A31 A21;A11,A22)	0.1123 (0.0010)
pCID(A21 A32;A11,A22)	0.0929 (0.0010)		
pCID(A22 A31;A11,A21)	0.0122 (0.3227)		
pCID(A22 A32;A11,A21)	0.0205 (0.1638)		
pCID(A11 A32;A21,A22,A31)	0.0123 (0.5465)		
pCID(A21 A32;A11,A22,A31)	<b>0.0553 (0.0020)</b>	pCID(A32 A21;A11,A22,A31)	0.0576 (0.0350)
pCID(A22 A32;A11,A21,A31)	0.0162 (0.5415)		
pCID(A31 A32;A11,A21,A22)	0.0298 (0.1788)		
CID(B A11)	0.0036 (0.9999)		
CID(B A21)	0.0202 (0.2468)		
CID(B A22)	0.0012 (0.9990)		
CID(B A31)	0.0137 (0.4905)		
CID(B A32)	0.0090 (0.6563)		

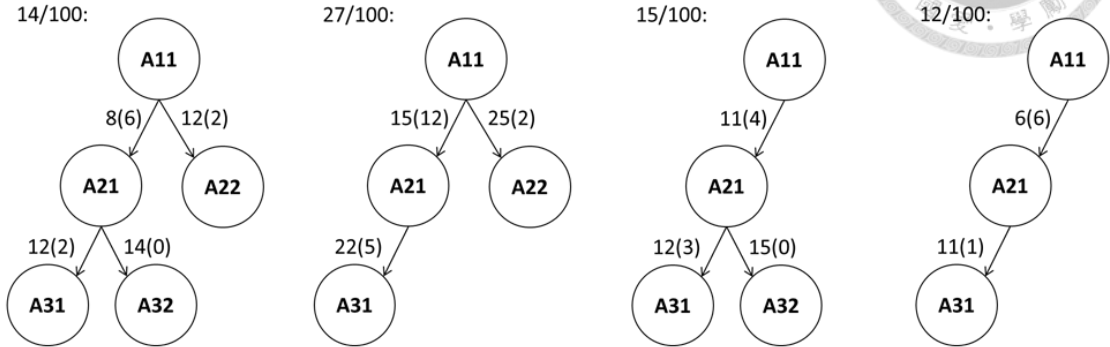
pCID(A31|A21; A11, A22) (value: 0.1131 > 0.1123), and pCID(A21|A32; A11, A22, A31) was more significant than pCID(A32|A21; A11, A22, A31) ( $p$ -value: 0.0020 < 0.0350) even though pCID(A21|A32; A11, A22, A31) value (0.0553) was smaller than pCID(A32|A21; A11, A22, A31) value (0.0576), respectively. When considering the negative-control node B as the source node, it had all insignificant values of CID at the first step of GRN inference and was isolated from the other nodes. Therefore, the resulting network was identical to our setting showing in Figure 4.4.

We also collected all networks reconstructed under the source node was A11 in the simulations for  $N = 25, 50$  and  $100$ ; networks consisting of the same set of nodes were grouped together and the groups occurred at least 5 times were shown in Figure 4.5. Fourteen resulting networks obtained the correct network structure among these one hundred simulations for  $N = 25$ , sixty-five correct networks were restructured for  $N = 50$  and eighty-one correct networks were for  $N = 100$ . For  $N = 25$ , 54% of the simulations only revealed the partial network; when using a larger sample ( $N = 50$ ), as few as 10 simulations obtained partial network; moreover, there were not any partial network under the sample of size  $N = 100$ . In addition, we could observe that the two nodes were sometimes discarded to produce the partial networks, if the proportion of gene expressions of the target gene actually determined by the expression levels of the source gene was lower than 76% (Figure 4.4) under the sample of size  $N = 25$ . In other words, the edges between (A11, A22) and (A21, A32) could be missed in the reconstruction of pseudo network. Similarly, the edge between (A11, A22) would be discarded when the proportion of A22 gene expressions actually determined by A11 was lower than 60% (Figure 4.4) under the sample of size  $N = 50$ . In this instance, the GRN would be accurately reconstructed in the large sample.

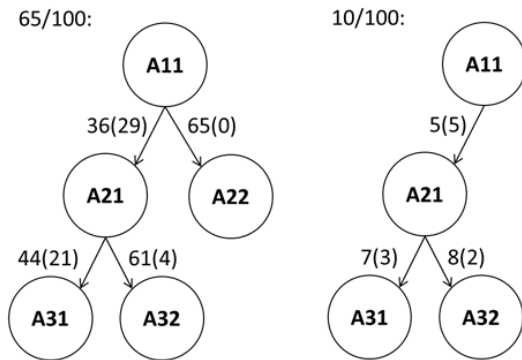
The asymmetric property of CID was utilized to infer causal effect in the network. When CID( $Y|X$ ) was more significant than CID( $X|Y$ ) or pCID( $Y|X; \mathbf{Z}$ ) was more significant than pCID( $X|Y; \mathbf{Z}$ ),  $Y$  was claimed to be the source of the relationship. In Figure 4.5 and Figure 4.6, the numbers of arrows which pointed to correct directions were shown beside the arrows outside of the parentheses whereas the numbers of incorrect directions in the parentheses. In Figure 4.6, we combined all the correct connections between two nodes from 100 simulations for  $N = 25, 50$  and  $100$ . When the sample of size  $N = 25$  and the source node was A11, there were 88% of networks to connect (A11, A21) together, 86% for (A21, A31), 55% for (A11, A22), and 40% for (A21, A32); 2% of the networks included the negative control node, B (Figure



**A. N = 25:**



**B. N = 50:**



**C. N = 100:**

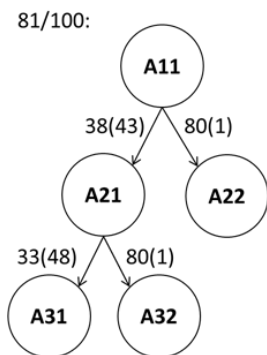


Figure 4.5: The results of the network reconstructed under the source node was A11 based on the procedure in Section 4.1 (Exclude the insignificant node by CID, and pick up the connected node which has the minimum significant CID/pCID  $p$ -value, if there existed at least two nodes which fitted the requests, we chose the node that had the maximum CID/pCID value) from 100 simulations of pseudo network for  $N = 25, 50$  and  $100$ , respectively. The numbers next to the arrows illustrate the number of connection from the source node to the target node; besides, the number of connection in the brackets illustrated the inverse direction.

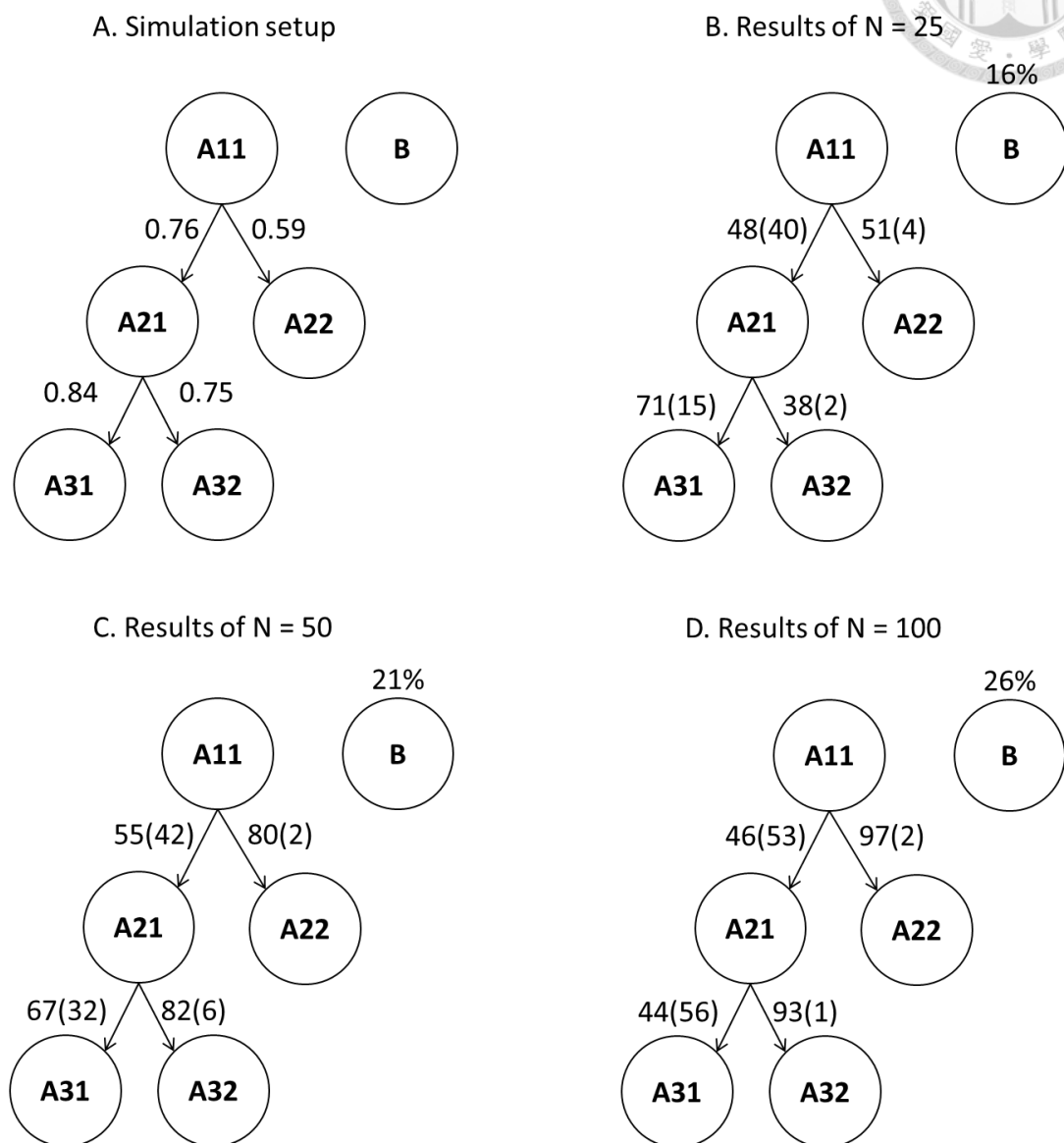


Figure 4.6: Pseudo network for the simulation study based on the procedure in Section 4.1 (Exclude the insignificant node by CID, and pick up the connected node which has the minimum significant CID/pCID  $p$ -value, if there existed at least two nodes which fitted the requests, we chose the node that had the maximum CID/pCID value). (A) The numbers next to the arrows illustrate the proportions of the objects in the sample that the expressions of the target node actually determined by the expressions of the source node. (B), (C) and (D) were the results which were combined with all connection from 100 simulations when the source node  $T_0$  was A11 for  $N = 25, 50$  and  $100$ , respectively.

4.6 B). When  $N = 50$ , 97%, 99%, 82%, and 88% of the networks contained the edges between (A11, A21), (A21, A31), (A11, A22) and (A21, A32), respectively, while 7% of them had the negative control node, B (Figure 4.6 C). When  $N = 100$ , 99%, 100%, 99%, and 94% of the networks contained the edges between (A11, A21), (A21, A31), (A11, A22) and (A21, A32), respectively, while 12% of them had the negative control node, B (Figure 4.6 D). When the negative control node, B, was set to be the source gene, 16% (Figure 4.6 B), 21% (Figure 4.6 C) and 26% (Figure 4.6 D) of the networks were significant build at  $\alpha = 0.05$ . However, the false networks were built spontaneously without consensus. All false networks started from B of the same combination of nodes only appeared less than or equal to five times in 100 simulations for  $N = 25, 50$  and 100. Therefore, CID/pCID method robustly identified the relationships between nodes and extended the association network.

The medians and interquartile ranges of some CID and pCID values summarized from 100 simulations were shown in Table 4.2. The CID values of A11 to a directed or undirected associated node were much larger than the CID values of A11 to the irrelevant node B. Also, it could be observed that  $CID(A11|A21) > CID(A11|A22)$ ,  $CID(A11|A31) > CID(A11|A32)$ , and  $CID(A11|A21)$  was larger than the maximum of  $CID(A11|A31)$  and  $CID(A11|A32)$  values. Therefore, CID value can not only distinguish the existence of association but also reflect the strength of the association and successfully pick the direct (or strongest) association among all possible connections. In addition, 100% of  $CID(A11|A21)$  and  $CID(A21|A11)$  values were declared significant if setting  $\alpha = 0.05$ . The pCID values further assisted to select next A11-related or A21-related node after eliminating the effects from A21 and A11, respectively. Among these pCID values, 100% of  $pCID(A21|A31; A11)$  values were significant at  $\alpha = 0.05$  and the medians of  $pCID(A21|A31; A11)$  values in different sample of size  $N$  were maximum, A31 was the most likely to be selected as A21-related node after eliminating the effects from A11. Furthermore, A22 was possibly picked up to connect with A11 based on 63% significance for the sample of size  $N = 25$  and 100% significance for  $N = 100$ ; A32 was possibly picked up to connect with A21 according to 97% significance for  $N = 50$ . In the final step, the chance A32 being selected in the elongation process to connect with A21 was only 29% for the sample of size  $N = 25$ , but there was 100% for  $N = 100$ ; the chance A22 being selected in the elongation process to connect with A11 was 83% for  $N = 50$ . On the other hand, the false positive rates of gene selection using either CID or pCID were all about 0.05.

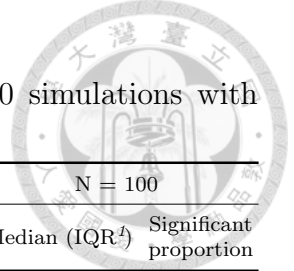


Table 4.2: Summary of the estimated CID/pCID values in 100 simulations with sample size  $N = 25, 50$  and  $100$ .

	N = 25		N = 50		N = 100	
	Median (IQR) <sup>†</sup>	Significant proportion	Median (IQR) <sup>†</sup>	Significant proportion	Median (IQR) <sup>†</sup>	Significant proportion
CID(A11 A21)	0.1967 (0.0534)	1.00	0.2049 (0.0527)	1.00	0.2319 (0.0378)	1.00
CID(A11 A22)	0.1100 (0.0568)	0.86	0.1232 (0.0522)	1.00	0.1402 (0.0331)	1.00
CID(A11 A31)	0.1348 (0.0631)	0.93	0.1457 (0.0610)	1.00	0.1600 (0.0345)	1.00
CID(A11 A32)	0.1130 (0.0708)	0.86	0.1233 (0.0499)	1.00	0.1328 (0.0377)	1.00
CID(A11 B)	0.0281 (0.0369)	0.06	0.0157 (0.0166)	0.13	0.0119 (0.0077)	0.16
CID(A21 A11)	0.1941 (0.0609)	1.00	0.2024 (0.0510)	1.00	0.2310 (0.0302)	1.00
pCID(A11 A22;A21)	0.0781 (0.0425)	0.74	0.0824 (0.0496)	0.96	0.0842 (0.0304)	1.00
pCID(A11 A31;A21)	0.0359 (0.0320)	0.22	0.0297 (0.0226)	0.55	0.0172 (0.0165)	0.83
pCID(A11 A32;A21)	0.0309 (0.0319)	0.19	0.0221 (0.0212)	0.40	0.0122 (0.0156)	0.72
pCID(A21 A22;A11)	0.0358 (0.0312)	0.19	0.0210 (0.0221)	0.33	0.0091 (0.0140)	0.61
pCID(A21 A31;A11)	0.1301 (0.0431)	1.00	0.1285 (0.0356)	1.00	0.1320 (0.0272)	1.00
pCID(A21 A32;A11)	0.0937 (0.0412)	0.93	0.1017 (0.0350)	1.00	0.0989 (0.0259)	1.00
pCID(A31 A21;A11)	0.1274 (0.0570)	0.92	0.1258 (0.0431)	1.00	0.1397 (0.0215)	1.00
pCID(A11 A22;A21,A31)	0.0764 (0.0536)	0.63	0.0772 (0.0461)	0.88	0.0838 (0.0385)	1.00
pCID(A11 A32;A21,A31)	0.0239 (0.0238)	0.04	0.0156 (0.0182)	0.09	0.0086 (0.0148)	0.23
pCID(A21 A22;A11,A31)	0.0202 (0.0242)	0.11	0.0126 (0.0197)	0.15	0.0009 (0.0156)	0.33
pCID(A21 A32;A11,A31)	0.0517 (0.0381)	0.52	0.0567 (0.0265)	0.97	0.0611 (0.0247)	1.00
pCID(A31 A22;A11,A21)	0.0160 (0.0211)	0.03	0.0057 (0.0137)	0.04	-0.0039 (0.0134)	0.07
pCID(A31 A32;A11,A21)	0.0295 (0.0273)	0.16	0.0237 (0.0238)	0.32	0.0195 (0.0181)	0.68
pCID(A22 A11;A21,A31)	0.0615 (0.0440)	0.18			0.0611 (0.0238)	0.86
pCID(A32 A21;A11,A31)			0.0486 (0.0222)	0.41		
pCID(A11 A32;A21,A22,A31)	0.0206 (0.0205)	0.01			0.0095 (0.0104)	0.14
pCID(A21 A32;A11,A22,A31)	0.0479 (0.0379)	0.29			0.0584 (0.0238)	1.00
pCID(A22 A32;A11,A21,A31)	0.0237 (0.0211)	0.01			0.0128 (0.0130)	0.02
pCID(A31 A32;A11,A21,A22)	0.0316 (0.0262)	0.08			0.0259 (0.0150)	0.41
pCID(A32 A21;A11,A22,A31)	0.0407 (0.0369)	0.02			0.0493 (0.0171)	0.59
pCID(A11 A22;A21,A31,A32)			0.0793 (0.0446)	0.83		
pCID(A21 A22;A11,A31,A32)			0.0123 (0.0189)	0.03		
pCID(A31 A22;A11,A21,A32)			0.0119 (0.0188)	0.07		
pCID(A32 A22;A11,A21,A31)			0.0143 (0.0192)	0.02		
pCID(A22 A11;A21,A31,A32)			0.0626 (0.0341)	0.35		
CID(B A11)	0.0273 (0.0285)	0.08	0.0167 (0.0163)	0.07	0.0119 (0.0100)	0.10
CID(B A21)	0.0220 (0.0231)	0.06	0.0144 (0.0129)	0.04	0.0103 (0.0072)	0.08
CID(B A22)	0.0187 (0.0222)	0.03	0.0114 (0.0117)	0.05	0.0075 (0.0060)	0.04
CID(B A31)	0.0199 (0.0239)	0.08	0.0125 (0.0149)	0.08	0.0079 (0.0086)	0.11
CID(B A32)	0.0188 (0.0158)	0.05	0.0131 (0.0171)	0.11	0.0078 (0.0064)	0.09

<sup>†</sup> IQR = interquartile range.

### 4.3 *Arabidopsis* microarray data analysis

C-repeat binding factors (CBF) would bind to the promoter regions of downstream cold-regulated (COR) genes and induce COR genes expression under cold stress (Thomashow *et al.*, 2001; McKhann *et al.*, 2008; Zhang *et al.*, 2013). We exercised the gene regulation network (GRN) inference on the expression dataset of *Arabidopsis Thaliana* under cold stress to reconstruct the well-known CBF-COR regulatory network. The detailed description about this dataset from TAIR database was in Section 3.3. After normalized and log2-transformed, the expressions of eight probes, three C-repeat binding factors (*CBF1* (probe ID: 254074\_at), *CBF2* (probe ID: 254075\_at) and *CBF3* (probe ID: 254066\_at)) and five COR gene family (*COR6.6* (probe ID: 246481\_s\_at), *COR78* (probe ID: 248337\_at), *COR47* (probe ID: 259570\_at), *COR15A* (probe ID: 263497\_at) and *COR15B* (probe ID: 263495\_at)), were taken to construct the GRN by CID/pCID method.

Three CBF genes took turns being the source of the regulation path elongation while the other probes were all considered as potential targets. Figure 4.7 (B), (C) and (D) showed the reconstructed pathways from the source CBF genes (rectangle nodes), respectively. The blue nodes and arrows denoted the CBF genes and the connections between CBF genes; the orange nodes and arrows denoted the COR genes and the connections between COR genes; the pink arrows denoted the connections between CBF and COR genes. The reconstructed pathways starting from *CBF2* (Figure 4.7 (C)) and *CBF3* (Figure 4.7 (D)) were the same; the pathway from the source gene *CBF1* (Figure 4.7 (B)) was similar to them and just the directions between CBF genes were different. Then we combined these pathways to reconstruct GRN in Figure 4.7 (A). Both *CBF1* and *CBF3* connected with *CBF2* in the sample, while *CBF3* had direct contact with the studied downstream COR genes. The *COR6.6* was the first receiver of the information passed down from CBF genes, which further influenced *COR78* and *COR15B*. By contrast, *COR47* and *COR15A* served as signal providers to the resulting path.

The heatmap and cluster analysis of CBF and COR relative gene expressions of different stressed conditions to their corresponding control samples was shown in Figure 4.8. The expressions of CBF genes on cold stress were increase early than COR genes, hence they would be the upstream of COR genes. Among them, *CBF3* had high expressions from 3hr to 12hr and lasted out longer than the other CBF genes. For that reason, *CBF3* might induce COR genes principally in our

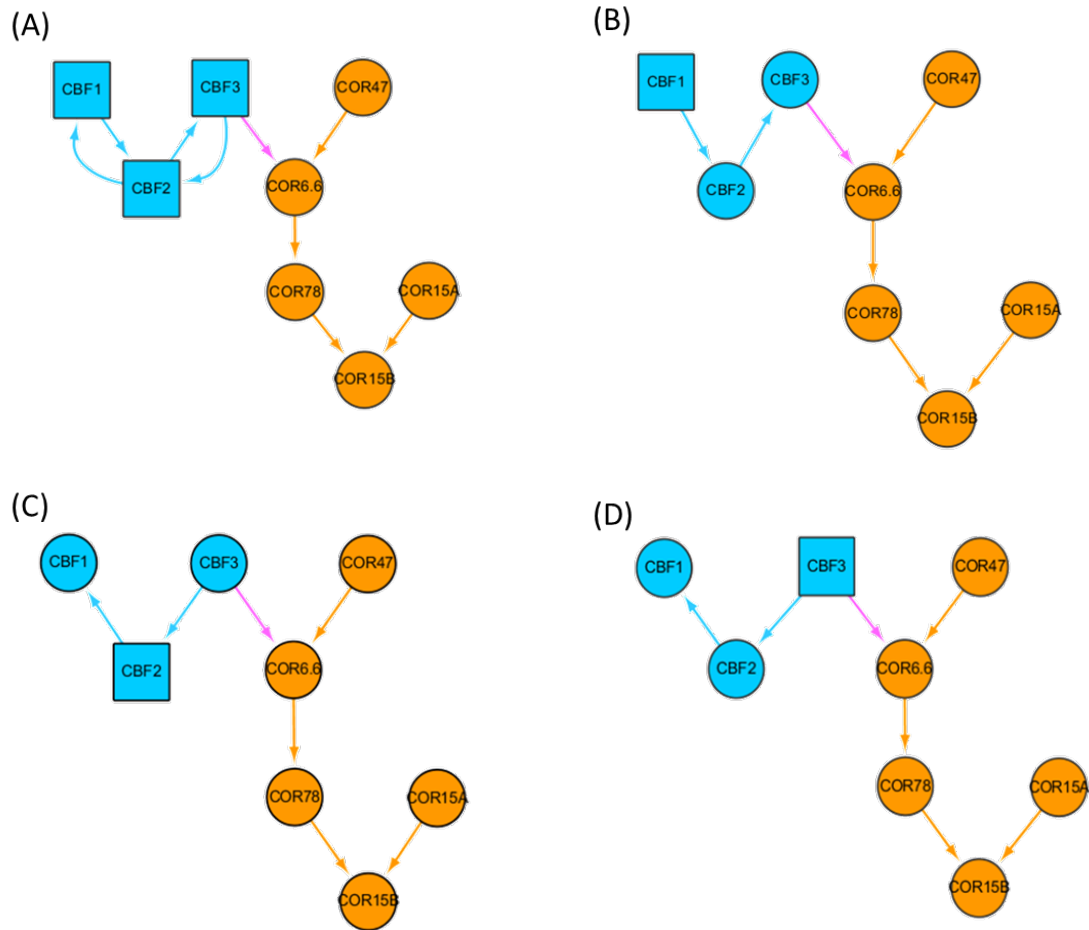


Figure 4.7: Reconstruction of CBF-COR regulatory network with eight genes under cold stress was based on CID/pCID method. (A) Combination of the pathways from three source genes (*CBF1*, *CBF2* and *CBF3*). (B), (C) and (D) were the pathways from the source genes, *CBF1*, *CBF2* and *CBF3*, respectively. Rectangle nodes indicate the source genes. Ellipse nodes are the candidate target genes.



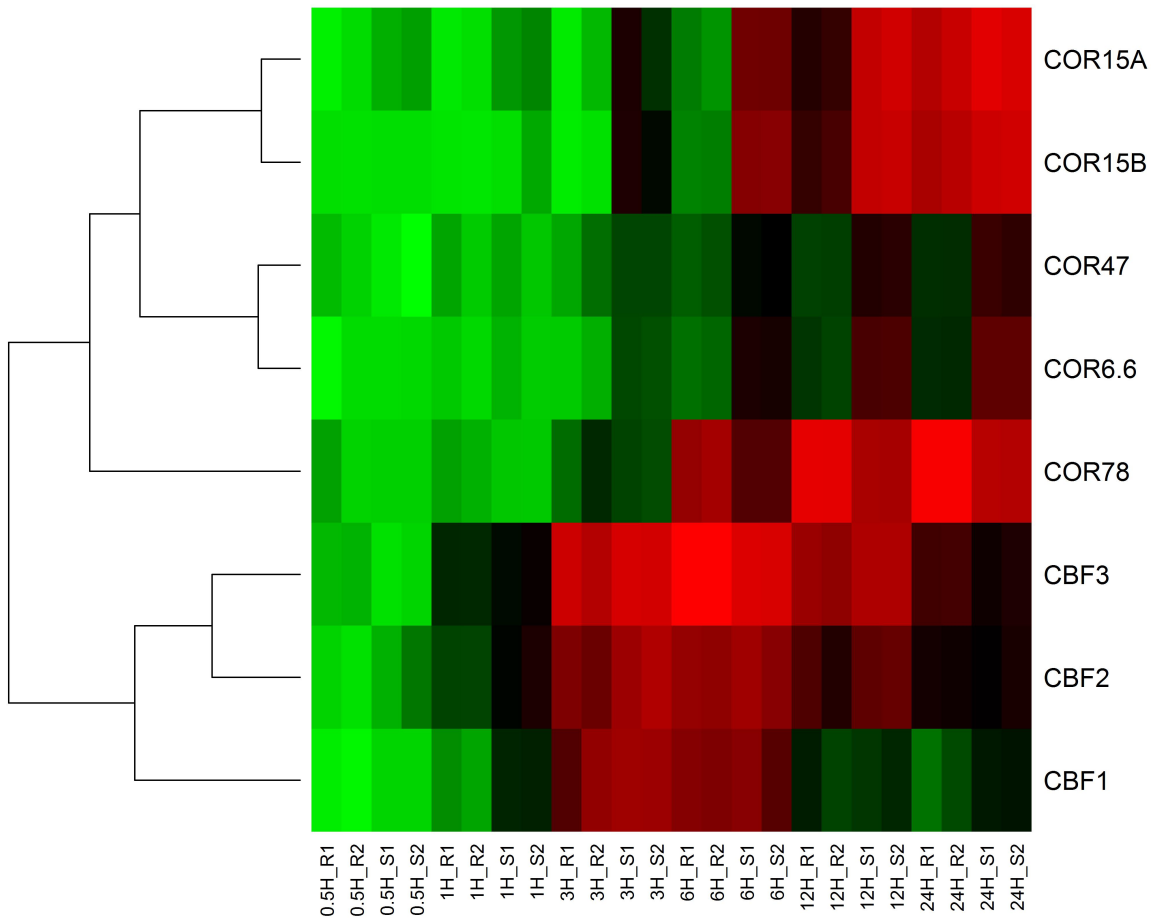
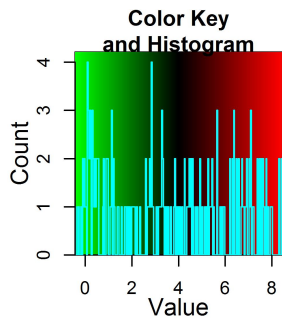


Figure 4.8: Cluster analysis and heatmap. A heatmap visualization of the log<sub>2</sub> relative treatment gene expression levels for the CBF and COR probes. R, root; S, shoot.

CID/pCID network results (Figure 4.7 (A)). Besides, expression of *COR47*, *COR78*, *COR15A*, *COR15B* and *COR6.6* was activated by *CBF3* in cold stress (Sakuma et al, 2006). On the other hand, *COR47* and *COR6.6* had similar expression levels; *COR15A* expressions were close to *COR15B*. The expressions of *COR78*, *COR15A* and *COR15B* had a tendency towards high level as time and *COR78* expressions occurred early of them. About the result of cluster analysis was shown the CBF and COR gene expressions could be separated into two groups.

Suppose that the regulation of CBF and COR genes was not discovered in biology. Each of eight probes was interchanged to be the source node of the gene pathway and the other seven probes would be the candidate target genes. The pathways of the CBF genes had exhibited in Figure 4.7 (B), (C) and (D). The other pathways of COR genes were shown in Figure 4.9 (B), (C), (D), (E) and (F). The reconstructed pathways starting from *COR15A* (Figure 4.9 (E)) and *COR15B* (Figure 4.9 (F)) were the same; the pathway from the source gene *COR47* (Figure 4.9 (B)) was similar to the result of *COR6.6* (Figure 4.9 (C)) and just the direction between *CBF1* and *CBF2* was different; the pathway from *COR78* (Figure 4.9 (D)) was different from others. However, there existed reverse direction between CBF and COR genes (pink arrows) in the pathways starting from each of COR genes. Based on the above pathways, the reconstructed GRN in Figure 4.9 (A) had 9% (5/54) reverse directions. Therefore, the reconstructed GRN based on CID/pCID could be more accurate while the source node had evidenced to be the upstream regulatory gene in biology.

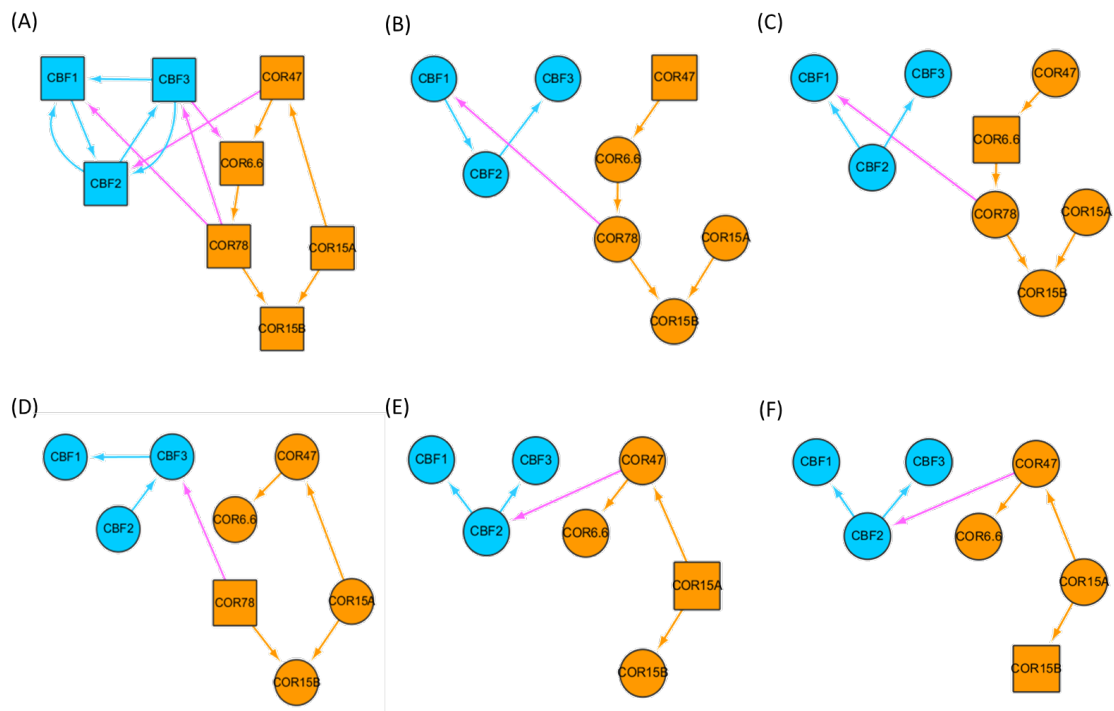
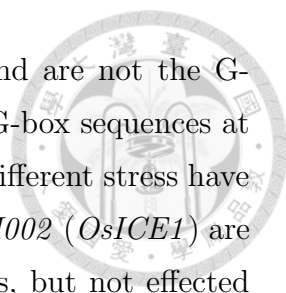


Figure 4.9: Reconstruction of CBF-COR regulatory network with eight genes under cold stress was based on CID/pCID method. (A) Combination of the pathways from all source genes (three CBF and five COR genes). (B), (C), (D), (E) and (F) were the pathways from the source genes, *COR47*, *COR6.6*, *COR78*, *COR15A* and *COR15B*, respectively. Rectangle nodes indicate the source genes. Ellipse nodes are the candidate target genes.

## 4.4 Rice microarray data analysis

The second dataset was to study the bHLH (basic helix-loop-helix) Pathway in rice (*Oryza Sativa*). The expressions data were downloaded from the NCBI-GEO database [ <http://www.ncbi.nlm.nih.gov/gds>] (accession numbers GSE6901 and GSE14275). The GSE6901 dataset includes gene expression of the 7-day-old light-grown rice seedlings under drought, salt and cold stresses from 9 samples (three biological replicates of each stress) as well as the gene expression from the adjacent controlled conditions of 3 samples. The GSE14275 dataset includes gene expression of the 14-day-old light-grown rice seedlings under heat shock stress from 3 samples and the gene expression from the adjacent controlled conditions of 3 samples. Both datasets hybridized the RNA samples on Affymetrix microarrays (NCBI-GEO accession number GPL2025). The raw expression data of 51,279 probes from 18 samples also went through pre-processing using the RMA method and log2 transformed. In this study, we were interested in the 167 genes that were previously reported as related genes involving in bHLH Pathway (Li *et al.*, 2006). Through matching the annotations of the affymetrix probe ID, we identified 128 bHLH-related probes in the microarray (Table B.1). Among them, 72 probes (61 genes) were called the G-box binders, which meant recognizing and binding to the G-box sequence (5'-CACGTG-3'), according to Li *et al.* (2006). We also downloaded the gene sequences of the bHLH-related genes in the microarray from RAP-DB (version 7.0) and found 104 probes (80 genes) containing G-box sequences in their promoter regions. The 72 probes recognize the G-box sequence and the 104 probes contain G-box sequences were designated as source and the candidate target genes, respectively, to construct the bHLH gene network. Besides, we match the 72 probes ID with 104 probes ID. There were 54 probes (45 genes) among these chosen probes to be appointed as source and the candidate target genes.

A family of transcription factors bHLH in plant plays principal role in developmental processes (Buck *et al.*, , 2003). The abiotic stresses affect the growth of crops. Up to the present, the functions of OsbHLH (*Oryza sativa* bHLH) transcription factors have not been studied completely. In this study, we explored the relationship of the OsbHLH gene expressions under the abiotic stresses by CID/pCID and the result of bHLH gene network was shown in Figure 4.10. The arrows indicate the association between two OsbHLH probes by CID/pCID. Rectangle nodes indicate the OsbHLH probes are the G-box binders and exclude G-box sequences. Ellipse



nodes indicate the *Os*bHLH probes include G-box sequences and are not the G-box binders. Octagon nodes are the G-box binders and include G-box sequences at the same time. The gray nodes represent that could respond in different stress have been verified in rice studies. *Os*bHLH001 (*Os*ICE2) and *Os*bHLH002 (*Os*ICE1) are induced at the protein level in response to cold and salt stresses, but not effected by cold stress on mRNA level (Nakamura *et al.*, 2011). *Os*bHLH006 (*RERJ1*) was shown to be up-regulated on drought stress (Kiribuchi *et al.*, 2005, Miyamoto *et al.*, 2013); *Os*bHLH009 (*Os*MYC) corresponded to Arabidopsis *At*MYC2 (Zhu *et al.*, 2005) and *At*MYC2 could induce the expression under drought stress (Abe *et al.*, 1997); *Os*bHLH062 (*Os*bHLH1) could be able to enhance the cold tolerance (Wang *et al.*, 2003); *Os*bHLH148 was induced by salt stress and resulted in activation under cold stress (Seo *et al.*, 2011); *Os*bHLH152 (*Os*PILI1) could reduce internode elongation under drought stress (Todaka *et al.*, 2012). Besides, *Os*bHLH001, *Os*bHLH002 and *Os*bHLH003 are related to the GO term, response to stress (GO: 0006950), from agriGO (GO Analysis Toolkit and Database for Agricultural Community). In Figure 4.10, we could observe that *Os*bHLH009 and *Os*bHLH148 connected with the downstream gene , *Os*bHLH006, respectively. Furthermore, *Os*bHLH006, *Os*bHLH009 and *Os*bHLH148 are important in drought stress.

In addition, *Os*bHLH010, *Os*bHLH024-1 (*Os*.10316.1.S1\_at), *Os*bHLH024-2 (*Os*.26054.1.S1\_s\_at), *Os*bHLH025-1 (*Os*.32770.1.S1\_x\_at), *Os*bHLH031, *Os*bHLH032, *Os*bHLH033-2 (*Os*.8796.2.S1\_a\_at), *Os*bHLH044, *Os*bHLH058, *Os*bHLH060, *Os*bHLH061, *Os*bHLH088, *Os*bHLH093, *Os*bHLH104-1 (*Os*.15089.1.S1\_at) and *Os*bHLH 104-2 (*Os*.44516.1.S1\_x\_at) might be the key roles in abiotic stresses because they had a lot of connections within these genes and with the other *Os*bHLH probes.

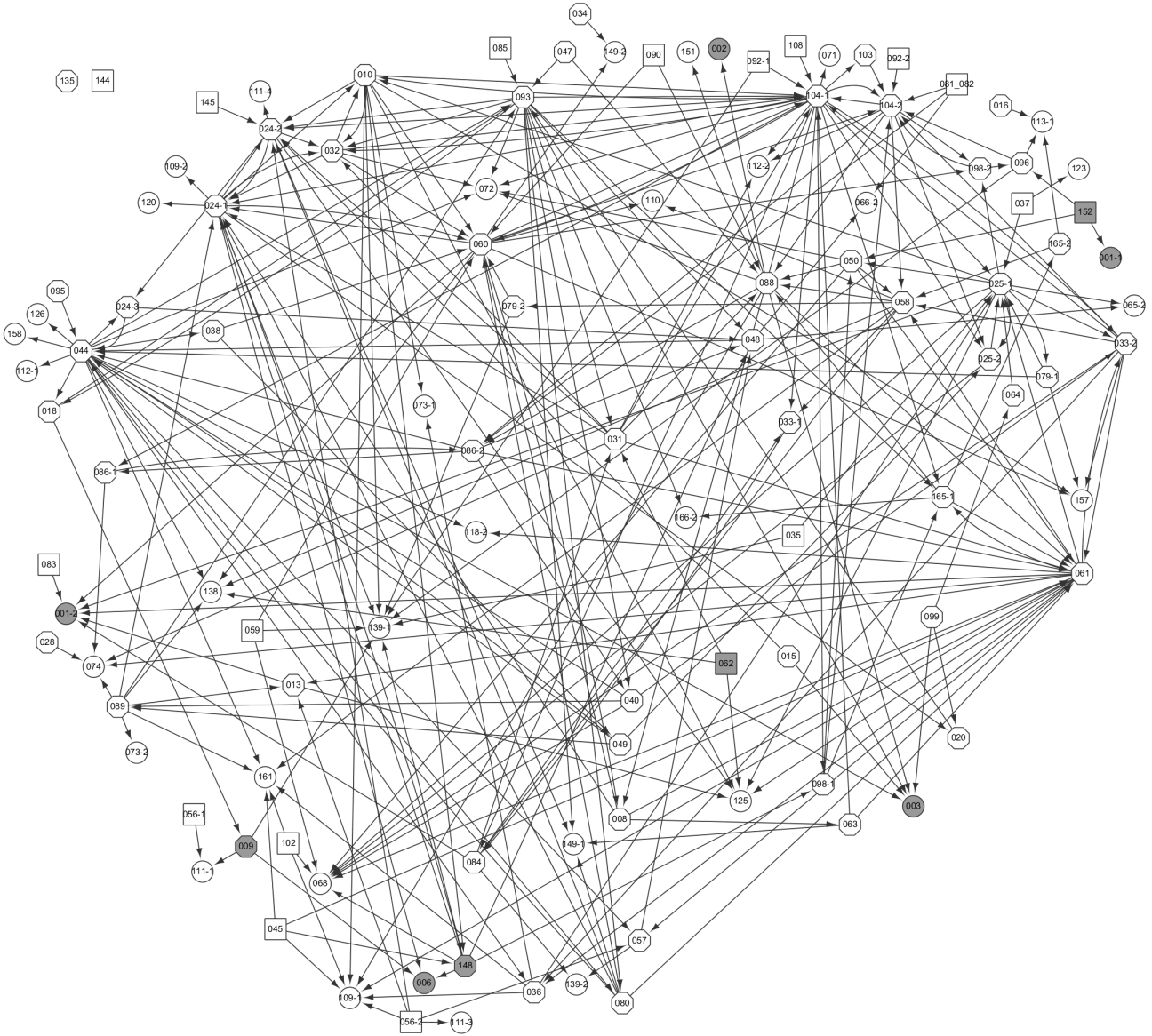



Figure 4.10: The gene regulatory network for OsbHLH rice seedlings contained the G-box binders and sequences under abiotic stresses is constructed by CID/pCID method from the NCBI-GEO database. Each node is the code of the OsbHLH number, for example 152 means the OsbHLH152. An arrow between nodes indicates a connection is determined by CID/pCID. Gray nodes show the genes are related to abiotic stresses have been confirmed from paper or GO term. Rectangle nodes indicate the OsbHLH probes are the G-box binders and exclude G-box sequences. Ellipse nodes indicate the OsbHLH probes include G-box sequences and are not the G-box binders. Octagon nodes are the G-box binders and include G-box sequences at the same time.

## 4.5 Discussion



For diminishing the computation of the programming, some irrelevant candidate target genes were eliminated in the first step of our proposed heuristic approach and were not proceed in the next steps. However, we use the same approach without eliminating the irrelevant genes to select the next genes for constructing the network. In order to compare the results with these two programmings, we use the same 100 simulations of pseudo network for sample size  $N = 25, 50$  and  $100$ . Consider a particular simulation with  $N = 50$ , which is the same as that is used in Table 4.1, the CID and pCID values as well as their  $p$ -values are shown in Table 4.3. Starting from the source node, A11, the first selected node is A22 and the direction is set from A11 to A22. For proceeding the steps, the results are  $A11 \rightarrow A21$ ,  $A21 \rightarrow A31$ ,  $A21 \rightarrow A32$  and  $A21 \rightarrow B$ . Next starting from the other source node, B, there are all insignificant values of CID at the first step of GRN inference and was isolated from the other nodes. Hence, the resulting network is distinct from the pseudo network in Figure 4.4. We obtain another connection,  $A21 \rightarrow B$ , which is unsuitable for our expectations.

We also collect all networks reconstructed under the source node is A11 in the simulations for  $N = 25, 50$  and  $100$ ; networks consisting of the same set of nodes are grouped together and the groups occur at least 5 times are shown in Figure 4.11. Fifteen resulting networks match the correct network structure among these one hundred simulations for  $N = 25$ , thirty-eight correct networks are restructured for  $N = 50$  and forty-seven correct networks are for  $N = 100$ . However, these proportions of correct networks with different sample sizes are almost less than the results of our proposed heuristic approach in Figure 4.5. Because of using the new approach may increase additional connections besides the complete network. There are 23% and 39% of the simulations have additional connections with the negative-control node B for  $N = 50$  and  $100$ , respectively. In addition, there also have the partial networks. For  $N = 25$ , 47% of the simulations only reveal the partial network; when using a larger sample ( $N = 50$ ), as few as 8 simulations obtain partial network; moreover, there were not any partial network under the sample of size  $N = 100$ .

In Figure 4.12, we combine all the correct connections between two nodes from 100 simulations for  $N = 25, 50$  and  $100$ . When the sample of size  $N = 25$  and the source node is A11, there are 88% of networks to connect (A11, A21) together, 92% for (A21, A31), 57% for (A11, A22), and 44% for (A21, A32); 14% of the networks



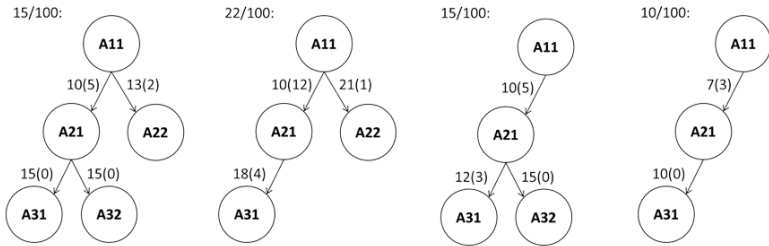
Table 4.3: The estimated CID and pCID values in one of the 100 simulations with sample size  $N = 50$ .

CID/pCID	Estimate ( $p$ -value)	CID/pCID	Estimate ( $p$ -value)
CID(A11 A21)	0.1936 (0.0010)		
CID(A11 A22)	<b>0.2028 (0.0010)</b>	CID(A22 A11)	0.1791 (0.0010)
CID(A11 A31)	0.1612 (0.0010)		
CID(A11 A32)	0.1281 (0.0010)		
CID(A11 B)	0.0129 (0.4136)		
pCID(A11 A21;A22)	<b>0.1013 (0.0010)</b>	PCID(A21 A11;A22)	0.0934 (0.0010)
pCID(A11 A31;A22)	0.0639 (0.0020)		
pCID(A11 A32;A22)	0.0534 (0.0010)		
pCID(A11 B;A22)	-0.0040 (0.5894)		
pCID(A22 A21;A11)	0.0582 (0.0060)		
pCID(A22 A31;A11)	0.0446 (0.0100)		
pCID(A22 A32;A11)	0.0500 (0.0090)		
pCID(A22 B;A11)	-0.0182 (0.9860)		
pCID(A11 A31;A21,A22)	0.0097 (0.2208)		
pCID(A11 A32;A21,A22)	0.0130 (0.1858)		
pCID(A11 B;A21,A22)	-0.0068 (0.7642)		
pCID(A21 A31;A11,A22)	<b>0.1131 (0.0010)</b>	pCID(A31 A21;A11,A22)	0.1123 (0.0010)
pCID(A21 A32;A11,A22)	0.0929 (0.0010)		
pCID(A21 B;A11,A22)	0.0063 (0.5994)		
pCID(A22 A31;A11,A21)	0.0122 (0.3227)		
pCID(A22 A32;A11,A21)	0.0205 (0.1638)		
pCID(A22 B;A11,A21)	-0.0150 (0.9950)		
pCID(A11 A32;A21,A22,A31)	0.0123 (0.5465)		
pCID(A11 B;A21,A22,A31)	0.0075 (0.6853)		
pCID(A21 A32;A11,A22,A31)	<b>0.0553 (0.0020)</b>	pCID(A32 A21;A11,A22,A31)	0.0576 (0.0350)
pCID(A21 B;A11,A22,A31)	0.0073 (0.6424)		
pCID(A22 A32;A11,A21,A31)	0.0162 (0.5415)		
pCID(A22 B;A11,A21,A31)	-0.0003 (0.9830)		
pCID(A31 A32;A11,A21,A22)	0.0298 (0.1788)		
pCID(A31 B;A11,A21,A22)	0.0194 (0.4486)		
pCID(A11 B;A21,A22,A31,A32)	0.0149 (0.5854)		
pCID(A21 B;A11,A22,A31,A32)	<b>0.0327 (0.0410)</b>		
pCID(A22 B;A11,A21,A31,A32)	0.0032 (0.9840)		
pCID(A31 B;A11,A21,A22,A32)	0.0254 (0.5754)		
pCID(A32 B;A11,A21,A22,A31)	0.0484 (0.0609)		
CID(B A11)	0.0036 (0.9999)		
CID(B A21)	0.0202 (0.2468)		
CID(B A22)	0.0012 (0.9990)		
CID(B A31)	0.0137 (0.4905)		
CID(B A32)	0.0090 (0.6563)		

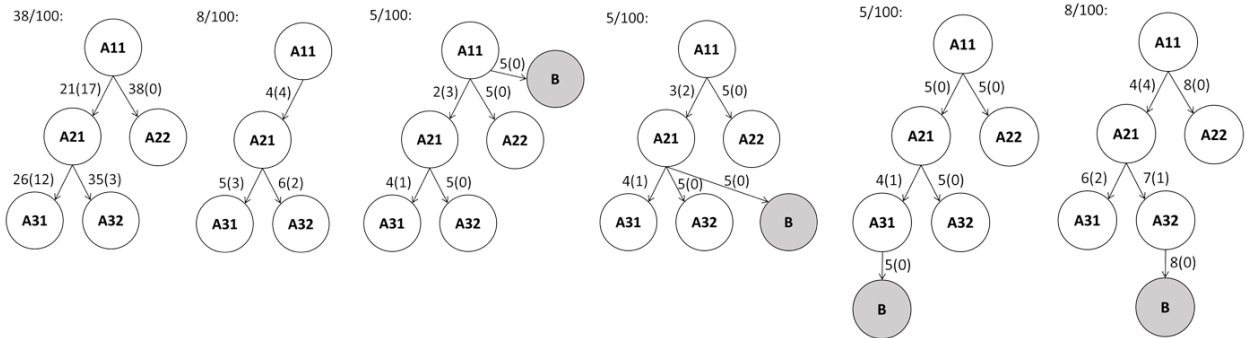




**A.  $N = 25$ :**



**B.  $N = 50$ :**



**C.  $N = 100$ :**

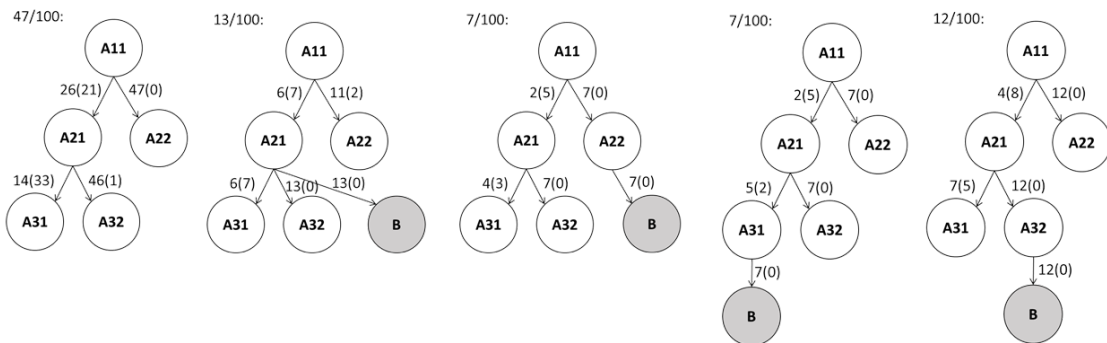


Figure 4.11: The results of the network reconstructed from 100 simulations of pseudo network for  $N = 25, 50$  and  $100$ , respectively. The numbers next to the arrows illustrate the number of connection from the source node to the target node; besides, the number of connection in the brackets illustrated the inverse direction.

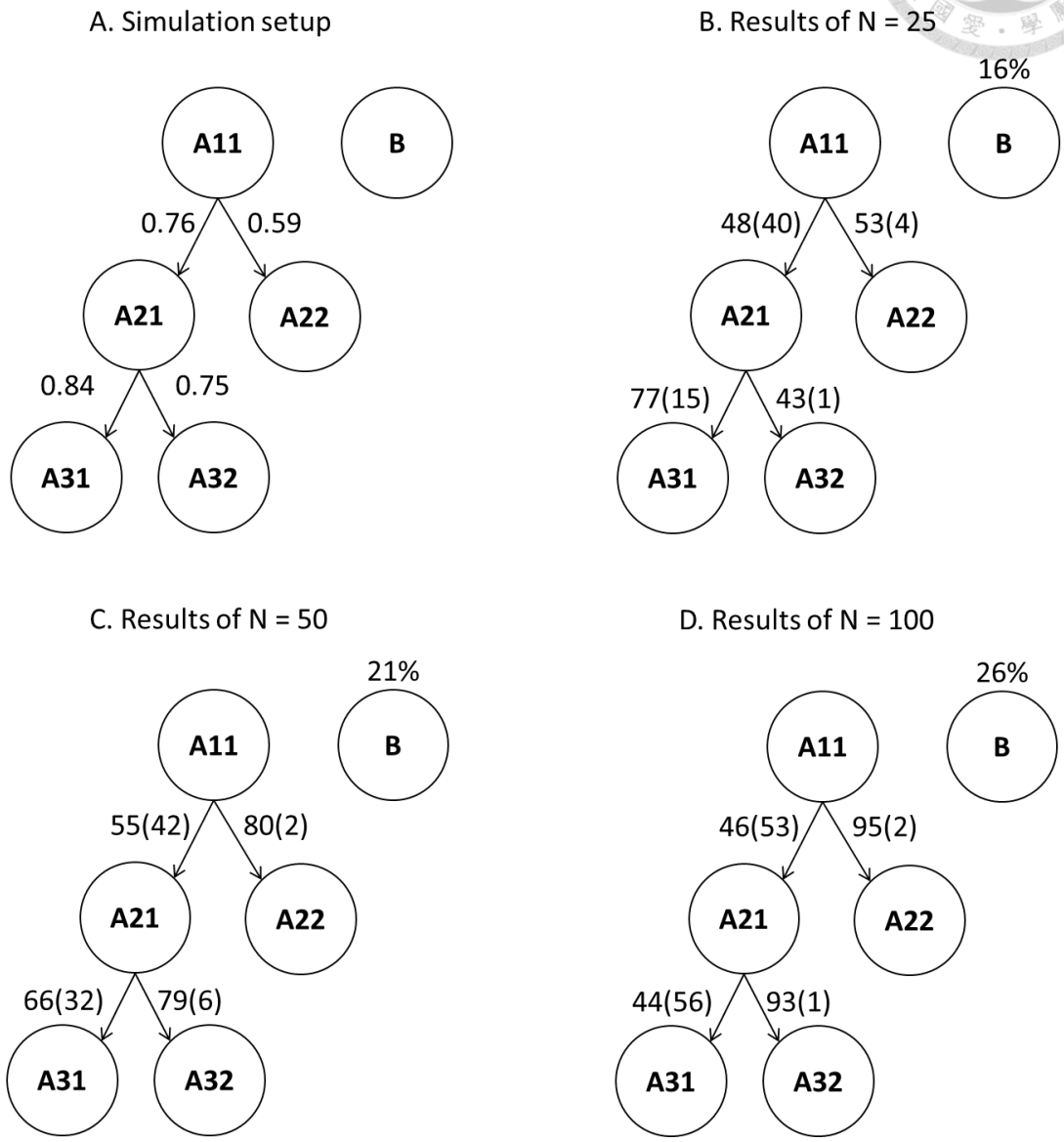


Figure 4.12: Pseudo network for the simulation study based on the procedure (Pick up the connected node which has the minimum significant CID/pCID  $p$ -value, if there existed at least two nodes which fitted the requests, we chose the node that had the maximum CID/pCID value). (A) The numbers next to the arrows illustrate the proportions of the objects in the sample that the expressions of the target node actually determined by the expressions of the source node. (B), (C) and (D) were the results which were combined with all connection from 100 simulations when the source node  $T_0$  was A11 for  $N = 25, 50$  and  $100$ , respectively.

include the negative control node B (Figure 4.12 B). When  $N = 50$ , 97%, 98%, 82%, and 85% of the networks contain the edges between (A11, A21), (A21, A31), (A11, A22) and (A21, A32), respectively, while 46% of them had node B (Figure 4.12 C). When  $N = 100$ , 99%, 100%, 97%, and 94% of the networks contain the edges between (A11, A21), (A21, A31), (A11, A22) and (A21, A32), respectively, while 48% of them had node B (Figure 4.12 D). We can observe that the proportions of networks which are combined all correct edges are similar to the outcomes in Figure 4.6. However, the proportions of networks include node B are larger than the results of our proposed approach and go up as the sample size increases. On the other source node B, 16% (Figure 4.12 B), 21% (Figure 4.12 C) and 26% (Figure 4.12 D) of the networks are significant build at  $\alpha = 0.05$ . All false networks start from B of the same combination of nodes only appear less than or equal to five times in 100 simulations for  $N = 25, 50$  and 100. Therefore, our proposed heuristic approach which was eliminated some irrelevant nodes in the first step based on CID has more accuracy.



## Chapter 5

# Conclusions

We have proposed a strategy to select explanatory variables that are relevant to the target variable using the CID along with the pCID without interference from other essential variables. The proposed method is more sensitive to curvilinearity and more specific to linearity than the PCC/pPCC method. It is also demonstrated in the simulations that the proposed procedure is able to quantify various types of associations in a stepwise manner. It also had the potential to index different levels of curvilinearity. While practicing on real microarray data, we have noticed that the CID/pCID procedure can not only identify cold-responsive genes but can also capture sample-specific gene-gene interactions. Biologists may find the proposed strategy useful in their efforts to extract meaningful relationships among genes out of the noise when meta analysis is of large interest in the post-genomic era.

In addition, we have extended the CID/pCID method to construct the gene regulatory network. The proposed heuristic approach can obtain more accurate reconstructed network when the sample size increase in the simulation study. While exercising a known gene regulatory network inference on gene expression data, we have observed that the CID/pCID programming can acquire more consistent pathway if the source gene is an upstream gene which has evidenced in biology. On the other hand, we practice an unknown gene regulatory network inference to supply not only some notable genes but also the new network. Biologists can verify the gene-gene interactions according to the experiments and explore the biological properties.



## References

Abe, H., Yamaguchi-Shinozaki, K., Urao, T., Iwasaki, T., Hosokawa, D., and Shinozaki, K. (1997), "Role of Arabidopsis MYC and MYB homologs in drought- and abscisic acid-regulated gene expression." *The Plant Cell*, 9: 1859-1868.

Akhtar, M., Jaiswal, A., Taj, G., Jaiswal, J. P., Qureshi, M. I., and Singh, N. K. (2012), "DREB1/CBF transcription factors: their structure, function and role in abiotic stress tolerance in plants." *Journal of Genetics*, 91: 385-395.

Baba, K., Shibata, R., and Sibuya, M. (2004), "Partial correlation and conditional correlation as measures of conditional independence." *Australian and New Zealand Journal of Statistics*, 46(4): 657-664.

Benjamini, Y., and Hochberg, Y. (1995), "Controlling the false discovery rate: a practical and powerful approach to multiple testing." *Journal of the Royal Statistical Society. Series B. Methodological*, 57: 289-300.

Buck, M.J., and Atchley, W.R. (2003), "Phylogenetic analysis of plant basic helix-loop-helix proteins." *Journal of Molecular Evolution*, 56: 742-750.

Du, Z., Zhou, X., Ling, Y., Zhang, Z., and Su, Z. (2010), "agriGO: a GO analysis toolkit for the agricultural community." *Nucleic Acids Research*, 38: W64-W70.

Faith, J.J., Hayete, B., Thaden, J.T., Mogno, I., Wierzbowski, J., Cottarel, G., Kasif, S., Collins, J.J. and Gardner, and T.S. (2007), "Large-scale mapping and validation of escherichia coli transcriptional regulation from a compendium of expression profiles." *PLOS Biology*, 5(1): 54-66.

Fowler, S., and Thomashow, M. F. (2002), "Arabidopsis transcriptome profiling indicates that multiple regulatory pathways are activated during cold acclimation in addition to the CBF cold response pathway." *The Plant Cell*, 14: 1675-1690.

Friedman, J.H. (1991), "Multivariate adaptive regression splines." *The Annals of*



*Statistics*, 19: 1-67.

Fuente, A. de la, Bing, N., Hoeschele, I., and Mendes, P. (2004), "Discovery of meaningful associations in genomic data using partial correlation coefficients." *BMC Bioinformatics*, 20(18): 3565-3574.

Gentleman, R. C., Carey, V. J., Bates, D. M., Bolstad, B., Dettling, M., Dudoit, S., Ellis, B., Gautier, L., Ge, Y., Gentry, J., Hornik, K., Hothorn, T., Huber, W., Iacus, S., Irizarry, R., Leisch, F., Li, C., Maechler, M., Rossini, A. J., Sawitzki, G., Smith, C., Smyth, G., Tierney, L., Yang, J. Y., and Zhang, J. (2004), "Bioconductor: open software development for computational biology and bioinformatics." *Genome Biology*, 5: R80.

Gilmour, S.J., Fowler, S.G., and Thomashow, M.F. (2004), "Arabidopsis transcriptional activators CBF1, CBF2, and CBF3 have matching functional activities." *Plant Molecular Biology*, 54(5): 767-781.

Hayfield, T., and Racine, J. S. (2008), "Nonparametric econometrics: the np package." *Journal of Statistical Software*, 27(5).

Hsing, T., Liu, L.Y., Brun, M., and Dougherty, E.R. (2005), "The coefficient of intrinsic dependence (feature selection using el CID)." *Pattern Recognition*, 38(5): 623-636.

Huala, E., Dickerman, A. W., Garcia-Hernandez, M., Weems, D., Reiser, L., LaFond, F., Hanley, D., Kiphart, D., Zhuang, M., Huang, W., Mueller, L. A., Bhattacharyya, D., Bhaya, D., Sobral, B.W., Beavis, W., Meinke, D.W., Town, C. D., Somerville, C., Rhee, and S. Y. (2001), "The Arabidopsis information resource (TAIR): a comprehensive database and web-based information retrieval, analysis, and visualization system for a model plant." *Nucleic Acids Research*, 29: 102-105.

Irizarry, R. A., Hobbs, B., Collin, F., BeazerBarclay, Y. D., Antonellis, K. J., Scherf, U., Speed, and T. P. (2003), "Exploration, normalization, and summaries of high density oligonucleotide array probe level data." *Biostatistics*, 4: 249-264.

Jain, M. (2012), "Next-generation sequencing technologies for gene expression profiling in plants." *Briefings in Functional Genomics*, 11(1): 63-70.

Kim, S. (2012), "ppcor: partial and semi-partial (part) correlation." [Http://CRAN.R-project.org/package=ppcor](http://CRAN.R-project.org/package=ppcor).

Kiribuchi, K., Jikumaru, Y., Kaku, H., Minami, E., Hasegawa, M., Kodama, O., Seto, H., Okada, K., Nojiri, H., and Yamane, H. (2005), "Involvement of the basic helix-loop-helix transcription factor RERJ1 in wounding and drought stress responses in rice plants." *Biosci. Biotechnol. Biochem.*, 69(5): 1042-1044.

Krouk, G., Lingeman, J., Colon, A.M., Coruzzi, G., and Shasha, D. (2013), "Gene regulatory networks in plants: learning causality from time and perturbation." *Genome Biology*, 14: 123.

Lee, B.-h., Henderson, D. A., and Zhu, J.-K. (2005), "The Arabidopsis cold-responsive transcriptome and its regulation by ICE1." *The Plant Cell*, 17: 3155-3175.

Li, X., Duan, X., Jiang, H., Sun, Y., Tang, Y., Yuan, Z., Guo, J., Liang, W., Chen, L., Yin, J., Ma, H., Wang, J., and Zhang, D. (2006), "Genome-wide analysis of basic/helix-loop-helix transcription factor family in rice and Arabidopsis." *Plant Physiology*, 141(4): 1167-1184.

Liu, L.Y.D. (2005), *Coefficient of intrinsic dependence: a new Measure of association*, Ph.D. Dissertation, Texas A& M University: College Station, Texas, USA.

Liu, L.Y.D., Chang, L.Y., Kuo, W.H., Hwa, H.L., Shyu, M.K., Chang, K.J., and Hsieh, F.J. (2012), "In silico prediction for regulation of transcription factors on their shared target genes indicates relevant clinical implications in a breast cancer population." *Cancer Informatics*, 11: 113-137.

Liu, L.Y.D., Chen, C.Y., Chen, M.J.M., Tsai, M.S., Lee, C.H.S., Phang, T.L., Chang, L.Y., Kuo, W.H., Hwa, H.L., Lien, H.C., Jung, S.M., Lin, Y.S., Chang, J.K., and Hsieh, F.J. (2009), "Statistical identification of gene association by CID in application of constructing ER regulatory network." *BMC Bioinformatics*, 10: 85.

Liu, Q., Kasuga, M., Sakuma, Y., Abe, H., Miura, S., Yamaguchi-Shinozaki, K., and Shinozaki, K. (1998), "Two transcription factors, DREB1 and DREB2, with an EREBP/AP2 DNA binding domain separate two cellular signal transduction pathways in drought- and low-temperature-responsive gene expression, respectively, in Arabidopsis." *The Plant Cell*, 10(8): 1391-1406.

Mardis, E.R. (2008), "Next-generation DNA sequencing methods." *Genomics and Human Genetics*, 9: 387-402.

McKhann, H.I., Gery, C., Bérard, A., Lévêque, S., Zuther, E., Hincha, D.K., Mita, S.D., Brunel, D., and Téoulé, E. (2008), "Natural variation in CBF gene sequence, gene expression and freezing tolerance in the Versailles core collection of *Arabidopsis thaliana*." *BMC Plant Biology*, 8(1): 105.

Miyamoto, K., Shimizu, T., Mochizuki, S., Nishizawa, Y., Minami, E., Nojiri, H., Yamane, H., and Okada, K. (2013), "Stress-induced expression of the transcription factor RERJ1 is tightly regulated in response to jasmonic acid accumulation in rice." *Protoplasma*, 250(1): 241-249.

Nakamura, J., Yuasa, T., Huong, T.T., Harano, K., Tanaka, S., Iwata, T., Phan, T., Iwaya-Inoue, M. (2011) "Rice homologs of inducer of CBF expression (OsICE) are involved in cold acclimation." *Plant Biotechnology*, 28(3): 303-309.

Priness, I., Maimon, O., and Ben-Gal, I. (2007), "Evaluation of gene-expression clustering via mutual information distance measure." *BMC Bioinformatics*, 8:111.

Sakuma, Y., Maruyama, K., Osakabe, Y., Qin, F., Seki, M., Shinozaki, K., and Yamaguchi-Shinozaki, K. (2006), "Functional analysis of an *Arabidopsis* transcription factor, DREB2A, involved in drought-responsive gene expression." *Plant Cell*, 18: 1292-1309.

Schadt, E.E., Lamb, J., Yang, X., Zhu, J., Edwards, S., Guhathakurta, D., Sieberts, S.K., Monks, S., Reitman, M., Zhang, C., Lum, P.Y., Leonardson, A., Thieringer, R., Metzger, J.M., Yang, L., Castle, J., Zhu, H., Kash, S.F., Drake, T.A., Sachs, A., and Lusk, A.J. (2005), "An integrative genomics approach to infer causal associations between gene expression and disease." *Nature Genetics*, 37: 710-717.

Seo, J.S., Joo, J., Kim, M.J., Kim, Y.K., Nahm, B.H., Song, S.I., Cheong, J.J., Lee, J.S., Kim, J.K., and Choi, Y.D. (2011), "OsHLLH148, a basic helix-loop-helix protein, interacts with OsJAZ proteins in a jasmonate signaling pathway leading to drought tolerance in rice." *Plant J.*, 65(6): 907-921.

Shrinet, J., Jain, S., Jain, J., Bhatnagar, R.K., and Sunil, S. (2014), "Next generation sequencing reveals regulation of distinct aedes microRNAs during chikungunya virus development." *PLoS Neglected Tropical Diseases*, 8(1):e2616.

Suh, E.B., Dougherty, E.R., Kim, S., Bittner, M.L., Chen, Y., Russ, D.E., Martino, and R.L. (2003), "Parallel computation and visualization tools for codetermination



analysis of multivariate gene expression relations." *Computational and Statistical Approaches to Genomics*, 227-240.

Thomashow, M.F., Gilmour, S.J., Stockinger, E.J., Jaglo-Ottosen, K.R., and Zarka, D.G. (2001), "Role of the Arabidopsis CBF transcriptional activators in cold acclimation." *Physiologia Plantarum*, 112(2): 171-175.

Tittarelli, A., Santiago, M., Morales, A., Meisel, L. A., and Silva, H. (2009), "Isolation and functional characterization of cold-regulated promoters, by digitally identifying peach fruit cold-induced genes from a large EST dataset." *BMC Plant Biology*, 9: 121.

Todaka, D., Nakashima, K., Maruyama, K., Kidokoro, S., Osakabe, Y., Ito, Y., Matsukura, S., Fujita, Y., Yoshiwara, K., Ohme-Takagi, M., Kojima, M., Sakakibara, H., Shinozaki, K., and Yamaguchi-Shinozaki, K. (2012), "Rice phytochrome-interacting factor-like protein OsPIL1 functions as a key regulator of internode elongation and induces a morphological response to drought stress." *Proceedings of the National Academy of Sciences*, 109(39): 15947-15952.

Tsai, C.A., and Liu, L.Y.D. (2013), "Identifying gene set association enrichment using the coefficient of intrinsic dependence." *PLoS One*, 8(3): e58851.

Wang, Y., Zhang, Z., He, X., Zhou, H., Wen, Y., Dai, J., Zhang, J. and Chen, S. (2003), "A rice transcription factor OsbHLH1 is involved in cold stress response." *Theoretical and Applied Genetics*, 107: 1402-1409.

Wettenhall, J. M., Simpson, K. M., Satterley, K., and Smyth, G. K. (2006), "affymGUI: a graphical user interface for linear modeling of single channel microarray data." *Bioinformatics*, 22: 897-899.

Zhang, Q., Jiang, N., Wang, G-L., Hong, Y., and Wang, Z. (2013), "Advances in understanding cold sensing and the cold-responsive network in rice." *Advances in Crop Science and Technology*, 1: 104.

Zhu, Z.F., Sun, C.Q., Fu, Y.C., Qian, X.Y., Yang, J.S., Wang, X.K. (2005), "Isolation and analysis of a novel MYC gene from rice." *Journal of Genetics and Genomics*, 32(4): 393-398.



## Appendix A

# The inference of pseudo network

Suppose A11 and B were randomly generated from  $N(1, 1)$ . In the pair genes  $(S, T)$ , if  $S$  was expressed, the expression level of  $T$  was distributed as  $N(1, 0.25)$ ; otherwise, the expression level of  $T$  was distributed as  $N(-1, 0.25)$ . The critical value of these two distribution was setted at the mean value minus two standard deviations and which value was calculated to be zero. The binding efficiency ( $b$ ) for  $\{A11, A21\}$ ,  $\{A11, A22\}$ ,  $\{A21, A31\}$ , and  $\{A21, A32\}$  were 0.9, 0.7, 0.9, and 0.8, respectively. The approximate proportions of gene expressions of the target gene actually determined by the expression levels of the source gene were expressed as  $P(S \rightarrow T)$  and the inferences were shown as follows.

- $P(A11 > 0) \simeq 0.84$ .  
The binding efficiency  $b_{\{A11, A21\}}$  was 0.9.  
Therefore  $P(A11 \rightarrow A21) \simeq 0.84 \times 0.9 \simeq 0.76$ .
- $P(A11 > 0) \simeq 0.84$  and  $b_{\{A11, A22\}} = 0.7$ .  
Then  $P(A11 \rightarrow A22) \simeq 0.84 \times 0.7 \simeq 0.59$ .
- $P(A11 > 0) \simeq 0.84$  and  $b_{\{A21, A31\}} = 0.9$ .

$$\begin{aligned}
 P(A21 > 0) &= P[I_{(A11 \rightarrow A21)} N(A11, 0.25) > 0] + P[I_{(A11 \rightarrow A21)} N(-1, 0.25) > 0] \\
 &= b_{\{A11, A21\}} [P(0 < A11 < 1) P(N(0, 0.25) > -0.5) + P(A11 > 1)] \\
 &\quad + (1 - b_{\{A11, A21\}}) P(N(-1, 0.25) > 0) \\
 &\simeq 0.9 \times (0.34 \times 0.84 + 0.5) + 0.24 \times 0.025 \\
 &\simeq 0.713
 \end{aligned}$$

$$P(A11 \rightarrow A31) \simeq 0.713 \times 0.9 \simeq 0.64.$$

$$\text{Thus } P(A21 \rightarrow A31) \simeq \frac{0.64}{0.76} \simeq 0.84.$$

- $P(A_{21} > 0) \simeq 0.713$  and  $b_{\{A_{21}, A_{32}\}} = 0.8$ .

$$P(A_{11} \rightarrow A_{32}) \simeq 0.713 \times 0.8 \simeq 0.57.$$

$$\text{Thus } P(A_{21} \rightarrow A_{32}) \simeq \frac{0.57}{0.76} \simeq 0.75.$$





# Appendix B

## Supplement table

Table B.1: GenBank accession number of OsbHLH members is in this study.

OsbHLH number	GenBank accession number	Affymetrix probe ID	MSU ID	RAP ID
OsbHLH001-1 (OsICE2)	AK102594.1	Os.13595.1.S1_at	LOC_Os01g70310	Os01g0928000
OsbHLH001-2 (OsICE2)	BI796438	Os.13595.2.S1_x_at	LOC_Os01g70310	Os01g0928000
OsbHLH002 (OsICE1)	AK109915.1	Os.56356.1.S1_at	LOC_Os11g32100	Os11g0523700
OsbHLH003 (RAI1)	AK103779.1	Os.5860.1.S1_at	LOC_Os03g04310	Os03g0135700
OsbHLH004-1	AK063669.1	Os.46563.1.S1_at	LOC_Os10g39750	Os10g0544200
OsbHLH004-2	AK063669.1	Os.46563.1.S1_a_at	LOC_Os10g39750	Os10g0544200
OsbHLH005 (TDR)	AK106761.1	Os.50000.1.S1_at	LOC_Os02g02820	Os02g0120500
OsbHLH006 (RERJ1)	AB040744.1	Os.6043.1.S1_at	LOC_Os04g23550	Os04g0301500
OsbHLH008	AK064943.1	Os.3825.1.S1_at	LOC_Os01g13460	Os01g0235700
OsbHLH009 (OsMYC)	AY536428.1	Os.46443.1.S1_at	LOC_Os10g42430	Os10g0575000
OsbHLH010	AK064946.1	Os.46956.1.S1_at	LOC_Os01g50940	Os01g0705700
OsbHLH013 (OSB1/Ra)	AB021079.1	Os.2233.1.S1_at	LOC_Os04g47080	Os04g0557800
OsbHLH015	AK111704.1	Os.49810.1.S1_at	LOC_Os04g47040	Os04g0557200
OsbHLH016 (OSB2)	AB021080.1	Os.57542.1.S1_at	LOC_Os04g47059	Os04g0557500
OsbHLH018	AK120539.1	Os.7441.1.S1_at	LOC_Os03g51580	Os03g0725800
OsbHLH020	AK107190.1	Os.54959.1.S1_at	LOC_Os03g46860	Os03g0671800
OsbHLH024-1	AK106333.1	Os.10316.1.S1_at	LOC_Os01g39330	Os01g0575200
OsbHLH024-2	BM038927	Os.26054.1.S1_s_at	LOC_Os01g39330	Os01g0575200
OsbHLH024-3	BM038927	Os.26054.1.S1_at	LOC_Os01g39330	Os01g0575200
OsbHLH025-1	AK102964.1	Os.32770.1.S1_x_at	LOC_Os01g09990	Os01g0196300
OsbHLH025-2	AK102964.1	Os.32770.1.S1_at	LOC_Os01g09990	Os01g0196300
OsbHLH028	AK107675.1	Os.55212.1.S1_at	LOC_Os05g11070	Os05g0199800
OsbHLH031	AK100183.1	Os.5093.1.S1_at	LOC_Os08g38210	Os08g0490000
OsbHLH032	AK071315.1	Os.16741.1.S1_a_at	LOC_Os09g29930	Os09g0475400
OsbHLH033-1	AK072417.1	Os.8796.1.S2_s_at	LOC_Os01g65080	Os01g0871200
OsbHLH033-2	AK065024.1	Os.8796.2.S1_a_at	LOC_Os01g65080	Os01g0871200
OsbHLH034	AK068228.1	Os.52592.1.S1_at	LOC_Os02g49480	Os02g0726700
OsbHLH035	AK106292.1	Os.1443.1.S1_a_at	LOC_Os01g06640	Os01g0159800
OsbHLH036	AK110619.1	Os.56950.1.S1_at	LOC_Os05g07120	Os05g0163900
OsbHLH037	AK068593.1	Os.26488.1.S1_at	LOC_Os01g11910	Os01g0218100
OsbHLH038	AK109616.1	Os.56209.1.S1_at	LOC_Os08g33590	Os08g0432800
OsbHLH040	AK106649.1	Os.54743.1.S1_at	LOC_Os03g15440	Os03g0260600
OsbHLH044	AK107555.1	Os.31303.1.S1_at	LOC_Os03g08930	Os03g0188400
OsbHLH045	AK058809.1	Os.46600.1.S1_at	LOC_Os10g23050	Os10g0376900
OsbHLH047	AK107626.1	Os.55174.1.S1_at	LOC_Os08g37730	Os08g0483900
OsbHLH048	AK107898.1	Os.55338.1.S1_at	LOC_Os02g52190	Os02g0759000
OsbHLH049	AK060695.1	Os.51109.1.S1_at	LOC_Os02g46560	Os02g0691500
OsbHLH050	AK062895.1	Os.51474.1.S1_at	LOC_Os04g50090	Os04g0590800
OsbHLH056-1 (OsIRO2)	AK073385.1	Os.12498.1.S1_at	LOC_Os01g72370	Os01g0952800
OsbHLH056-2 (OsIRO2)	AK104991.2	Os.12498.2.S1_at	LOC_Os01g72370	Os01g0952800
OsbHLH057	AK068361.1	Os.26508.2.S1_a_at	LOC_Os07g35870	Os07g0543000
OsbHLH058	AK063498.1	Os.49628.1.S1_at	LOC_Os05g38140	Os05g0455400
OsbHLH059	AK103434.1	Os.17893.1.S1_at	LOC_Os02g02480	Os02g0116600
OsbHLH060	AK102951.1	Os.18333.1.S1_at	LOC_Os08g04390	Os08g0138500
OsbHLH061	AK068017.1	Os.27243.1.S1_at	LOC_Os11g38870	Os11g0601700
OsbHLH062 (OsbHLH1)	AY222337.1	Os.34549.1.S1_at	LOC_Os07g43530	Os07g0628500
OsbHLH063 (OsIRO3)	AK068704.1	Os.9216.1.S1_at	LOC_Os03g26210	Os03g0379300

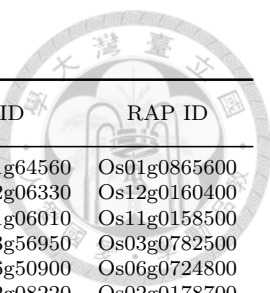
(Continued on next page)

(Continued from previous page)

Os bHLH number	GenBank accession number	Affymetrix probe ID	MSU ID	RAP ID
Os bHLH064	AK069790.1	Os.52897.1.S1_at	LOC_Os02g23823	Os02g0433600
Os bHLH065-1	AK059273.1	Os.6328.1.S1_at	LOC_Os04g41570	Os04g0493100
Os bHLH065-2	AK107304.1	Os.55009.1.S1_at	LOC_Os04g41570	Os04g0493100
Os bHLH066-1	AK072833.1	Os.51847.1.S1_x_at	LOC_Os03g55220	Os03g0759700
Os bHLH066-2	AK064057.1	Os.51847.2.S1_at	LOC_Os03g55220	Os03g0759700
Os bHLH068	AK069366.1	Os.25006.1.A1_at	LOC_Os04g53990	Os04g0631600
Os bHLH071	AK119493.1	Os.45859.1.S1_at	LOC_Os01g01600	Os01g0105700
Os bHLH072	AK072848.1	Os.8589.1.S1_at	LOC_Os02g17680	Os02g0276900
Os bHLH073-1	AK121917.1	Os.10063.1.S1_at	LOC_Os05g14010	Os05g0228400
Os bHLH073-2	AK107340.1	Os.10063.2.S1_at	LOC_Os05g14010	Os05g0228400
Os bHLH074	AK065732.1	Os.38009.1.S1_at	LOC_Os01g13000	Os01g0230200
Os bHLH075	AK109094.1	Os.55989.1.S1_at	LOC_Os04g47810	Os04g0565900
Os bHLH076	AK107063.1	Os.54904.1.S1_at	LOC_Os02g45010	Os02g0671300
Os bHLH079-1	AK119183.1	Os.7751.1.S1_at	LOC_Os02g47660	Os02g0705500
Os bHLH079-2	AK107038.1	Os.7751.2.S1_at	LOC_Os02g47660	Os02g0705500
Os bHLH080	AK059041.1	Os.14318.1.S1_at	LOC_Os08g42470	Os08g0536800
Os bHLH081_082 (Os bHLH081 & Os bHLH082)	AK066188.1	Os.35707.1.S1_at	LOC_Os09g33580	Os09g0510500
Os bHLH083	AK065864.1	Os.23082.1.S1_at	LOC_Os05g01256	Os05g0103000
Os bHLH084	CB631822	Os.24540.1.A1_at	LOC_Os03g51910	Os03g0728900
Os bHLH085	AK121418.1	Os.38400.1.S1_at	LOC_Os09g29830	Os09g0474100
Os bHLH086-1	AK101279.1	Os.47378.1.S1_s_at	LOC_Os06g16400	Os06g0275600
Os bHLH086-2	AK103853.1	Os.32526.1.S1_at	LOC_Os06g16400	Os06g0275600
Os bHLH088	AK068324.1	Os.52614.1.S1_at	LOC_Os03g12940	Os03g0232000
Os bHLH089	AK100177.1	Os.33544.1.S1_at	LOC_Os03g58830	Os03g0802900
Os bHLH090	AK101063.1	Os.5763.1.S1_at	LOC_Os01g68700	Os01g0915600
Os bHLH092-1	AK099291.1	Os.10830.1.S1_at	LOC_Os09g32510	Os09g0501600
Os bHLH092-2	AK059036.1	Os.20775.1.S1_at	LOC_Os09g32510	Os09g0501600
Os bHLH093	AK108605.1	Os.55703.1.S1_at	LOC_Os04g28280	Os04g0350700
Os bHLH095	AK070970.1	Os.4952.1.S1_at	LOC_Os06g41060	Os06g0613500
Os bHLH096 (OsPTH1)	AY238991.1	Os.8790.1.S1_a_at	LOC_Os06g09370	Os06g0193400
Os bHLH098-1	AK067446.1	Os.27522.2.S1_at	LOC_Os03g58330	Os03g0797600
Os bHLH098-2	AK068388.1	Os.27522.1.S1_x_at	LOC_Os03g58330	Os03g0797600
Os bHLH099	AK066623.1	Os.8344.1.S1_at	LOC_Os07g08440	Os07g0182200
Os bHLH101	AK106689.1	Os.4548.1.S1_at	LOC_Os04g52770	Os04g0618600
Os bHLH102 (OsBP-5)	AK066763.1	Os.11675.1.A1_at	LOC_Os12g41650	Os12g0610200
Os bHLH103	AK060505.1	Os.19229.1.S1_a_at	LOC_Os03g43810	Os03g0639300
Os bHLH104-1	AK060245.1	Os.15089.1.S1_at	LOC_Os07g05010	Os07g0143200
Os bHLH104-2	CF326413	Os.44516.1.S1_x_at	LOC_Os07g05010	Os07g0143200
Os bHLH108	D43106	Os.23257.1.A1_at	LOC_Os06g06900	Os06g0164400
Os bHLH109-1	AK068254.1	Os.12030.1.S1_at	LOC_Os01g67480	Os01g0900800
Os bHLH109-2	AK121411.1	Os.50489.1.S1_at	LOC_Os01g67480	Os01g0900800
Os bHLH110	AK110833.1	Os.49337.1.S1_at	LOC_Os02g39140	Os02g0603600
Os bHLH111-1	AK068039.1	Os.7694.1.S1_at	LOC_Os04g41229	Os04g0489600
Os bHLH111-2	AK062301.1	Os.51233.1.S1_at	LOC_Os04g41229	Os04g0489600
Os bHLH111-3	AF467735.1	Os.57535.1.S1_at	LOC_Os04g41229	Os04g0489600
Os bHLH111-4	AF467735.1	Os.57535.1.A1_at	LOC_Os04g41229	Os04g0489600
Os bHLH112-1	AK100106.1	Os.5311.1.S1_at	LOC_Os08g39630	Os08g0506700
Os bHLH112-2	AK120902.1	Os.20361.1.A1_at	LOC_Os08g39630	Os08g0506700
Os bHLH113-1	CB624216	Os.27587.1.S1_at	LOC_Os10g40740	Os10g0556200
Os bHLH113-2	CB624215	Os.46626.1.S1_x_at	LOC_Os10g40740	Os10g0556200
Os bHLH118-1	AK109307.1	Os.25546.1.S1_at	LOC_Os01g51140	Os01g0707500
Os bHLH118-2	AK100208.1	Os.32078.1.S1_at	LOC_Os01g51140	Os01g0707500
Os bHLH120	AK070458.1	Os.51063.1.S1_at	LOC_Os09g28210	Os09g0455300
Os bHLH123 (OsLAX/LAX1)	AB115668.1	Os.38423.1.S1_at	LOC_Os01g61480	Os01g0831000
Os bHLH125	AK108587.1	Os.30617.1.S1_at	LOC_Os01g02110	Os01g0111500
Os bHLH126	AK109662.1	Os.56232.1.S1_at	LOC_Os02g48060	Os02g0710300
Os bHLH135	AK108042.1	Os.55414.1.S1_at	LOC_Os12g40590	Os12g0597800
Os bHLH138	AK065674.1	Os.28061.1.S1_at	LOC_Os03g27390	Os03g0391700
Os bHLH139-1	AK107002.1	Os.49098.1.S1_x_at	LOC_Os02g21090	Os02g0315600
Os bHLH139-2	AK106848.1	Os.49098.2.S1_at	LOC_Os02g21090	Os02g0315600
Os bHLH140	AK101749.1	Os.54081.1.S1_at	LOC_Os03g39432	Os03g0591300
Os bHLH141 (EAT1)	AK119509.1	Os.49995.1.S1_at	LOC_Os04g51070	Os04g0599300
Os bHLH142	AK106850.1	Os.54828.1.S1_at	LOC_Os01g18870	Os01g0293100
Os bHLH144	AK108728.1	Os.30520.1.S1_at	LOC_Os04g35010	Os04g0429400
Os bHLH145	AK107268.1	Os.54995.1.S1_at	LOC_Os04g35000	Os04g0429300
Os bHLH148	AK071734.1	Os.7116.1.S1_at	LOC_Os03g53020	Os03g0741100
Os bHLH149-1	AK099677.1	Os.14287.1.S1_at	LOC_Os01g64560	Os01g0865600

(Continued on next page)

(Continued from previous page)



OsBHLH number	GenBank accession number	Affymetrix probe ID	MSU ID	RAP ID
OsBHLH149-2	AK099677.1	Os.14287.1.S1_a_at	LOC_Os01g64560	Os01g0865600
OsBHLH150	AK074015.1	Os.48567.1.S1_at	LOC_Os12g06330	Os12g0160400
OsBHLH151	AK106579.1	Os.31883.1.A1_at	LOC_Os11g06010	Os11g0158500
OsBHLH152 (OsPIL1/OsPIL13)	AK105637.1	Os.5178.1.A1_s_at	LOC_Os03g56950	Os03g0782500
OsBHLH155	AK063523.1	Os.11409.1.S1_at	LOC_Os06g50900	Os06g0724800
OsBHLH157	AK110943.1	Os.15780.1.S1_at	LOC_Os02g08220	Os02g0178700
OsBHLH158	AK058439.1	Os.50771.1.S1_at	LOC_Os06g44320	Os06g0653200
OsBHLH160	AU031410	Os.18660.1.S1_x_at	LOC_Os11g02054	Os11g0111800
OsBHLH161	AK062951.1	Os.51497.1.A1_s_at	LOC_Os12g02020	Os12g0111400
OsBHLH162	AK063202.1	Os.11231.1.S1_at	LOC_Os05g27090	Os05g0337200
OsBHLH165-1 (Rb)	U39866.1	Os.57500.1.S1_at	LOC_Os01g39580	Os01g0577300
OsBHLH165-2 (Rb)	U39866.1	Os.57500.1.S1_x_at	LOC_Os01g39580	Os01g0577300
OsBHLH166-1	AK073378.1	Os.53575.1.S1_at	LOC_Os03g21970	Os03g0338400
OsBHLH166-2	AK073378.1	Os.53575.1.S1_s_at	LOC_Os03g21970	Os03g0338400