Master of Science

A MUTUAL INFORMATION-BASED MULTIPLE LEVEL DISCRETIZATION NETWORK INFERENCE FROM TIME-SERIES GENE EXPRESSION PROFILES

The Graduate School
of the University of Ulsan

Department of Electrical, Electronic and Computer Engineering
Cao Tuan Anh

A MUTUAL INFORMATION-BASED MULTIPLE LEVEL DISCRETIZATION NETWORK INFERENCE FROM TIME-SERIES GENE EXPRESSION PROFILES

Supervisor: Prof. Yung-Keun Kwon

A Dissertation

Submitted to
the Graduate School of the University of Ulsan
In Partial Fulfillment of the Requirements
for the Degree of

Master of Science

by

Cao Tuan Anh

Department of Electrical, Electronic and Computer Engineering
Ulsan, Korea
November 2021

Cao Tuan Anh의 공학석사 학위논문을 인준함

심사위원장 정 진 호(인)

심사위원 권 영 근(인)

심사위원 김 종 면(인)

울산대학교 대학원
2021년 11월

Abstract

Discovering a genetic regulatory network (GRN) from time series gene expression data plays an essential role in the field of biomedical research. This is because transcriptional regulation is a fundamental molecular mechanism that is involved in almost every aspect of life, from homeostasis to development, from metabolism to behavior, from reaction to stimuli to disease progression. So that many methods have been proposed for inferring GRNs. Among the proposed methods, Boolean networks are widely used. Although the Boolean network models give good results to some extent and are capable of handling data noise, information loss is their main drawback due to the simplicity of data representation, and this leads to the achieved results still being far from optimal.

Thus, it is needed to develop an efficient method which can infer large networks with a reliable result in an acceptable run time. In this regard, we propose a new method namely the mutual information based on multiple level discretization network inference (MIDNI) from time-series gene expression profiles. For each gene in the input network, real-valued gene expression values are discretized into binary or ternary depending on its distribution before feeding to the reference algorithm.

We validated MIDNI with four well-known inference methods, DBN, MICRAT, MIBNI, and GENIE3, through extensive simulations on both the artificial discretized and the artificial real-valued gene expression datasets. Our results illustrated that MIDNI significantly outperformed them in terms of both structural and dynamics accuracies. This implies that MIDNI is an efficient tool to reconstruct the gene regulatory networks, particularly, more efficiently for complex and large networks.

Acknowledgments

I am thankful to University of Ulsan for giving me scholarship to carry out my Master's research in Korea. The scholarship program granted me the opportunity to not only complete my academic research that took place under the guidance of the professors, but to also learn the Korean language, history, and culture.

I would like to express my sincere gratitude to my research advisor, Professor Kwon, Yung-Keun, for his expert advice and friendly guidance throughout my Master's degree program at the Complex System Computing Lab, University of Ulsan. He gave me the opportunity of research under his supervision and is always ready to support me when needed. Especially during the execution of my experiments, his encouragement and guidance helped me to achieve my success.

In addition, I would also like to extend my gratitude to the other professors in my thesis committee, Professor Chung, Jin Ho, and Professor Kim, Jong Myon for dedicating their time to review this thesis. Their insightful comments were beneficial contributions to the completion of this work.

I am grateful to all my lab mates for their kindness and ever available assistantship in the learning process as well as in life. They helped me get familiar with life at the first time in Korea, and we also often shared, discussed, cooperated in research career. I also would like to acknowledge the faculty office, the academic, and technical staffs with endless support.

Lastly, I am very thankful to my loved ones, family, and friends for their continuous love, support, faith, and encouragement throughout my studies.

Contents

List of figures

List of abbreviations

GRN Genetic or Gene Regulatory Network

MIDNI Multiple level Discretization Network Inference

MIFS Mutual Information based Feature Selection

DBN Dynamic Bayesian network

MICRAT Maximal Information coefficient with Conditional Relative Average entropy and Time series mutual information

MIBNI Mutual Information based Boolean Network Inference

GENIE3 Gene Network Inference with Ensemble of trees

Chapter 1

Introduction

1.1. Motivation

Gene regulatory networks (GRNs) determine the ensemble of underlying interactions among genes that carries out their expression. The clarification of GRNs is extensively important to understand the functioning and pathology of organisms and it still remains one of the major challenges of systems biology. Recently, due to the advent of high-throughput technologies, computational approaches have been proposed to infer GRNs from the measurement of gene expressions in various conditions (time-series and steady state) using statistical inference or machine learning techniques. Although the calculation speed is improved significantly, alternative data representation methods, however, are still used to speed up more the calculation process and also give high inference results. Within the methods of data representation, the binary representation is widely applied and has achieved many critical results. It is known as the Boolean network. Many exist methods have been proposed applying the binary data representation such as Best-Fit, MIBNI, etc. While these exist network inference methods have reached some maturity, their performance on real datasets remains far from optimal and calls for the permanent improvement of both inference result as well as computational speed. This is my motivation of this work.

1.2. Problem Statement

To infer the GRNs, many methods used the Boolean network model which has only two levels of data discretization. “1” denotes the active state (or activator) of the gene value while “0” means the inactive one (or inhibitor). Although inferring results have been reliable with acceptable accuracy and the methods are fast and efficient to capture the dynamical interactions of genes, all Boolean network methods in some cases show the poor performance due to the simplicity (leading to high information loss) of the Boolean representation of the data. To overcome this, some methods [1-3] applied a higher order discretization for genes whose values are denoted by 1, 0, and -1 meaning ‘upregulated’, ‘no-change’, and ‘downregulated’. However, most of these methods always used three levels while discretizing the data. This sometime is not optimal since among genes, not all of them are suitable to be discretized into three states. In fact, some is more appropriate for binary discretization, others else are more applicable for three levels of discretization.

To minimize the information loss while keeping the attractive advantages of the Boolean network for the network reconstruction problem, we propose a novel algorithm namely the mutual information based on multiple level discretization network inference from gene expression profiles (MIDNI) which used both the binary and ternary discretization depending upon the distribution of gene values. In this work, we will present how to determine which genes should be discretized into binary and which should not, and the most important stage is the proposed algorithm to infer the network and it results comparison.

1.3. Existing Solutions

Attempts have been made by industry professionals and researchers alike to solve the problem of inferring GRNs. And through these attempts, several approaches have been proposed. Some of these approaches are included in the following:

- Data-driven methods
- Methods based on the correlation score
- Methods based on the mutual information
- Tree-based methods
- Bayesian methods

Identifying the set of regulatory genes for the input network is a crucial step for the further understanding biological system. This topic has expectedly caught the interest of many researchers, who have gone ahead to publish their findings. From these things we find a more efficient and flexible approach which shows a better performance on large datasets.

1.4. Research Objectives

Using the representative data for inferring GRNs gives the result not much worse than those of using the raw expression data. Moreover, this leads the inference speed many times faster. The K-mean algorithm is widely used due to simple implementation and stably clustering. However, how to automatically determine the number of clusters is a tough question that is still challenging for many researchers. Many determine the optimal cluster according to plotting the Elbow score. This approach is emotional and for some cases, it is too difficult to choose which one is the best among the two.

Hence, the first goal of this study is to find the metric based on which we are able to determine the number of clusters as the parameter of the K-mean algorithm. When the data representation is done, the inference process is considered. Here, the aim is that it is necessary to develop an algorithm to find regulatory candidates for each target gene by considering all the interactions between genes instead of just considering the interaction of gene pairs relation between the target and the other genes in the network like the previous approaches. This is important because it resembles real biological systems. Finally, the algorithm not only gives a good performance in terms of the structure prediction but also should give adequate results in terms of dynamics accuracy, which other algorithms rarely evaluate. This is also important because the different structures of the networks might produce the same underlying dynamics. In short, the algorithm, in addition, to accurately predicting the network structure, also reveals the dynamic interactions among genes in the network.

1.5. Thesis Outline

This thesis is composed of five chapters, outlined as follows:

Chapter 1, which is the current chapter, introduces the subject of the thesis. It states the problem and motivation towards solving the stated problem. It further outlines existing solutions to the problem and our objective for conducting this research

Chapter 2 presents an overview and review of literature on referring GRNs. In this chapter I discuss some of the existing approaches to infer GRNs and works related to inference of GRNs that have been carried out in the past.

Chapter 3 describes the proposed methodology namely the mutual information based on multiple level discretization network inference from gene expression profiles. It

gives a detailed-on data representation and how to find out a set of candidate regulatory genes for each target gene of the input network.

Chapter 4 reports on the experiments carried out, and the results realized from these experiments. In this chapter, we will see how the proposed model performs in comparison to existing baselines, for both terms of structural and dynamic accuracies.

Chapter 5 concludes this thesis with a summary of the work done in the course of the research and provides directions for future studies.

Chapter 2

Backgrounds

2.1. Overview of gene regulatory network inference

A gene (or genetic) regulatory network (GRN) is a collection of molecular regulators that interact with each other and with other substances in the cell to govern the gene expression levels of mRNA and proteins which, in turn, determine the function of the cell. GRN also play a central role in morphogenesis, the creation of body structures, which in turn is central to evolutionary developmental biology.

Gene regulatory network inference allows us to understand biological systems behaviors. After that disease characteristics can be revealed out and at the end of this process is that drug therapy will be produced to deal with epidemics. Hence, gene regulatory network reconstruction plays an important role in understanding biological systems.

It can be said that gene regulatory networks are powerful abstractions of biological systems. In the late 1990s, thanks to the advent of high-throughput measurement technologies in biology, reconstructing the structure of such networks becomes a central computational challenge in systems biology. Although considerable progress has been made in the last two decades, however, the problem is certainly not solved in its entirely. Hence, researchers keep working hard to come up with better algorithms that give more positive results to solve the problem thoroughly. In this study, we propose a new approach to infer the network from Time-series gene expression profiles.

2.2. Related Works

Many approaches using various computational models have been proposed for the network inference problem. A data-driven methods is a class of GRN reconstruction methods by estimating genes dependencies directly from the data. In this class, the correlation score is widely used to associate to a pair of vector-valued measurements. The weighted gene co expression network analysis (WGCNA) [4] is a well-known correlation score-based method that has proved consistently reliable and widely adopted. However, the correlation coefficient fails to capture more complex statistical dependencies (such as non-linear relations) between expression patterns. To resolve this limitation, the mutual information (MI) has been employed. MI is an efficient information theoretic score which frequently used in determination of the regulatory relations. The relevance network [5] is the simplest model which computes MI between all pairs of genes and infers the presence of a regulatory interaction when MI is larger than a given threshold. The context likelihood of relatedness (CLR) algorithm [6] is an extension of the relevance networks approach to replace the probability distributions with empirical distributions of all MI scores. CLR applies an adaptive background correction step to eliminate false correlations and indirect influences. Algorithm for the reconstruction of accurate cellular networks (ARACNE) [7] is another information-theoretic algorithm for the reverse engineering of transcriptional networks from microarray data. ARACNE defines an edge as an irreducible statistical dependency between gene expression profiles that cannot be explained as an artifact of other statistical dependencies in the network. Finally, MRNET [8] uses an effective information-theoretic technique for feature selection based on a maximum relevance/minimum redundancy criterion. Despite certain results, the computational complexity is the major limitation of the data-driven

approaches which leads to consuming a lot of computational resources as well as the running time. Another limitation is that most of them are applicable only to small-size networks. They are difficult to test or show the poor performance with large-size networks.

In order to reduce the complexity of computation and speed up the running time, real-valued expression data conversion of Boolean value is widely used in many methods and used that discretization data as an input for inferring systems. For example, the Best-Fit [9] is proposed to cope with the Consistency Problem and find one or all consistent Boolean networks that are relative to the given examples. Another example of binary data using [10] that present an inference approach for a Boolean Network model of a GRN from limited transcriptomic or proteomic time series data based on prior biological knowledge of connectivity, constraints on attractor structure and robust design. The method is able to illustrate that inference of a Boolean network from limited time series data with constraints on connectivity that explains the observed state transitions. Mutual information-based Boolean network inference (MIBNI) method [11] also uses the binarized expression data as an input for inferring regulatory networks. MIBNI firstly identifies a set of initial regulatory genes using mutual information-based feature selection, and then improves the dynamics prediction accuracy by iteratively swapping a pair of genes between sets of the selected regulatory genes and the other genes. A genetic algorithm-based Boolean network inference (GABNI) method [12] is an upgrade approach from MIBNI. GABNI applies genetic algorithm (GA) to search an optimal set of regulatory genes in a wider solution space when MIBNI fails to find an optimal solution in a small-scale inference problem. Although GABNI shows a good performance for the large-scale networks, however, it employed a limited representation model of regulatory

functions. In this regard, a novel genetic algorithm combined with a neural network for the Boolean network inference was proposed [13], where a neural network is used to represent the regulatory function instead of an incomplete Boolean truth table used in the GABNI. In brief, although inferring results have been reliable with acceptable accuracy and the methods are fast and efficient to capture the dynamical interactions of genes, all Boolean network methods in some cases show the poor performance due to the simplicity (information loss) of the Boolean representation of the data.

To minimize the information loss while keeping the attractive advantages of the Boolean network for the network reconstruction problem, we propose a novel algorithm namely the mutual information based on multiple level discretization network inference from gene expression profiles (MIDNI). The algorithm first discretizes gene expression data where the discretization level depends on the distribution of gene values. In this study, the discretization level is limited to two or three. This is because if the discretization level is greater than three, the computational cost will increase. Hence, in each network, some genes are discretized into three to avoid information loss. Other genes are discretized into two for the purpose of optimizing the computational speed and handling data noise. We use the validity index (more detail see literatures [14] [15] [16], [17]) to determine whether which genes should be discretized into two and which should be into three. It depends on the distribution of every single gene in the network.

Subsequently, discretization genes are feed to the mutual information – based feature selection (MIFS) method [11]. We used MIFS to approximate the multi-variate mutual information between a target gene and a set of candidate regulatory genes. We also used a SWAP subroutine, which is a greedy algorithm wherein a gene in the set of regulatory genes selected by MIFS is iteratively swapped with another gene in the

set of unselected genes. This loop of MIFS and SWAP was presented in our earlier Boolean network inference (MIBNI) algorithm [11]. In this work, we modified them to adapt multi-level discretized expression values. We validated MIDNI performance with both the artificial and the real gene expression datasets in comparison with two well-known methods namely MICRAT, DBN and MIBNI. The experiment shows that our method outperformance them in term of dynamics accuracies. Furthermore, MIDNI is able to infer a network with noisy datasets and gives a reliable result. This indicates that MIDNI is a dependable method in inferring regulatory network from gene expression datasets.

Chapter 3

Materials and Methodology

3.1. Materials

3.1.1. A discretization network model

In this study, we employed a discretization network model to represent gene regulatory networks. A discretized network is represented by a directed graph *G(V, A)* where $V = \{v_1, v_2, \ldots, v_n\}$ is a set of nodes, $A = \{(v_i, v_j)\} \subseteq V \times V$ is a set of interactions, and a state value of gene v at time t, $V(t)$, is represented by l discrete values $\{0, 1, 2, \ldots, l-1\}$. We note that it is a Boolean network if $l = 2$ for all genes in V. Consider a target node $v \in V$ which is regulated by k genes, $u_1, u_2, \ldots, u_k$ ($\forall u_i \in V$). Let E and E_i are the sets of discretized expression values of genes v and u_i, respectively. The value $v(t+1)$ is updated by a discrete function $f: E_1 \times E_2 \times \ldots \times E_k \rightarrow E$ of the values of k regulatory genes $u_1, u_2, \ldots, u_k$ at time t. Hence, the update scheme of v can be described as the following formula:

$$v(t+1) = f(u_1(t), u_2(t), \ldots, u_k(t)).$$

We note that the update time lag used in this study is one and the number of all possible functions with respect to f is $l^{\Pi_i l_i}$ where l and l_i the cardinalities of E and E_i, respectively.

3.1.2. The discretization network inference problem

The discretization network inference problem is a problem to infer both a set of interactions and a set of update functions from a time-series gene expression data. The

inference performance can be evaluated by comparing the trajectory generated by the inferred network and the observed time-series gene expression. Let $v'(t)$ the predicted value of the gene v at time t in the inferred discretization network. We define the gene-wise dynamics consistency $C(v, v')$ as the similarity between the discretization trajectories of the observed gene expression $v(t)$ and the estimated gene expression $v'(t)$, as follows:

$$C(v, v') = \frac{\sum_{t=p}^{T} I(v(t) = v'(t))}{T - p}$$

where T is the total number of time-steps, p is time lag = 1 and I(.) is an indicator function that returns 1 if the condition is true, otherwise 0. In addition, the comparison starts at $t = 2$ by the assumed update time lag in this work. We finally define the dynamics accuracy of an inferred network as the average of gene-wise dynamics over all genes as follows:

$$\text{Dynamic Accuracy} = \frac{\sum_{i=1}^{N} C(v_i, v'_i)}{N}$$

, where N is the number of genes.

3.1.3. **Structure performance metrics**

When the structure of a gold standard or correct network is known, we can further evaluate the inference performance with respect to the network structure. To this end, we used three measures, precision, recall and structural accuracy. Precision is the ratio of correctly inferred connections over the total number of predictions as follows:

$$Precision = \frac{TP}{TP + FP}$$

where TP (true positive) and FP (false positive) denote the numbers of correctly and incorrectly predicted connections, respectively. Recall is the ratio of true predicted connections over the total number of actual connections:

$$Recall = \frac{TP}{TP + FN}$$

where FN (false negative) means the number of non-inferred connections in $G(V, A)$. Structural accuracy is the ratio of correct predictions out of all predictions as follows:

$$Structural\ Accuracy = \frac{TP + TN}{TP + FP + FN + TN}$$

where TN (true negative) is the number of correct negative predictions.

3.1.4. **Mutual information**

Our method selects a set of regulatory variables for each target variable based on some concepts of information theory. First, the entropy $H(X)$ of a discrete random variable X is defined to measure the uncertainty of X as follows:

$$H(X) = -\sum_{x \in X} p(x) \log p(x)$$

In addition, the join entropy $H(X, Y)$ of two discrete random variables X and Y with a joint probability distribution $p(x, y)$ is defined as follows:

$$H(X, Y) = -\sum_{x \in X} \sum_{y \in Y} p(x, y) \log p(x, y)$$

Finally, we used the mutual information of two discrete variables as follows:

$$I(X;Y) = H(X) + H(Y) - H(X,Y)$$

The larger the mutual information is, the more the variables are dependent on each other.

3.2. Methodology

In this study, we proposed a method called MIDNI for a multiple level discretization network inference from time-series gene expression data. Figure 1 demonstrates the overall structure of it. A real-valued time-series gene expression dataset is given as an input, and it is converted into a gene expression dataset using the K-means discretization method [14]. It divides all expression values of each gene into l dicretized values where the discretization level l is determined according to the distribution of the gene expression values. In case that the discretization level is two, the converted values of genes are marked by 0 (low) and 1 (high). On the other hand, they are marked by 0 (low), 1 (normal) and 2 (high) when the discretization level is three. For each target gene v, whose entropy value is non-zero, the subroutine MIFS is used to select k genes in V that have the most informative variables with the gene v. Then, the subroutine SWAP is applied to improve the gene-wise dynamics consistency by swapping the same number of variables between S and $W \setminus S$. Here, S is a selected set of candidate regulatory genes achieved by MIFS, and W is a set of genes in the network. This procedure is repeated by increasing the number k until an optimal set S is found (the gene-wise dynamics consistency of the gene v has reached perfect and equals 1) or k reaches a user-parameter K which indicates the maximum number of regulatory genes for the gene v to be inferred.

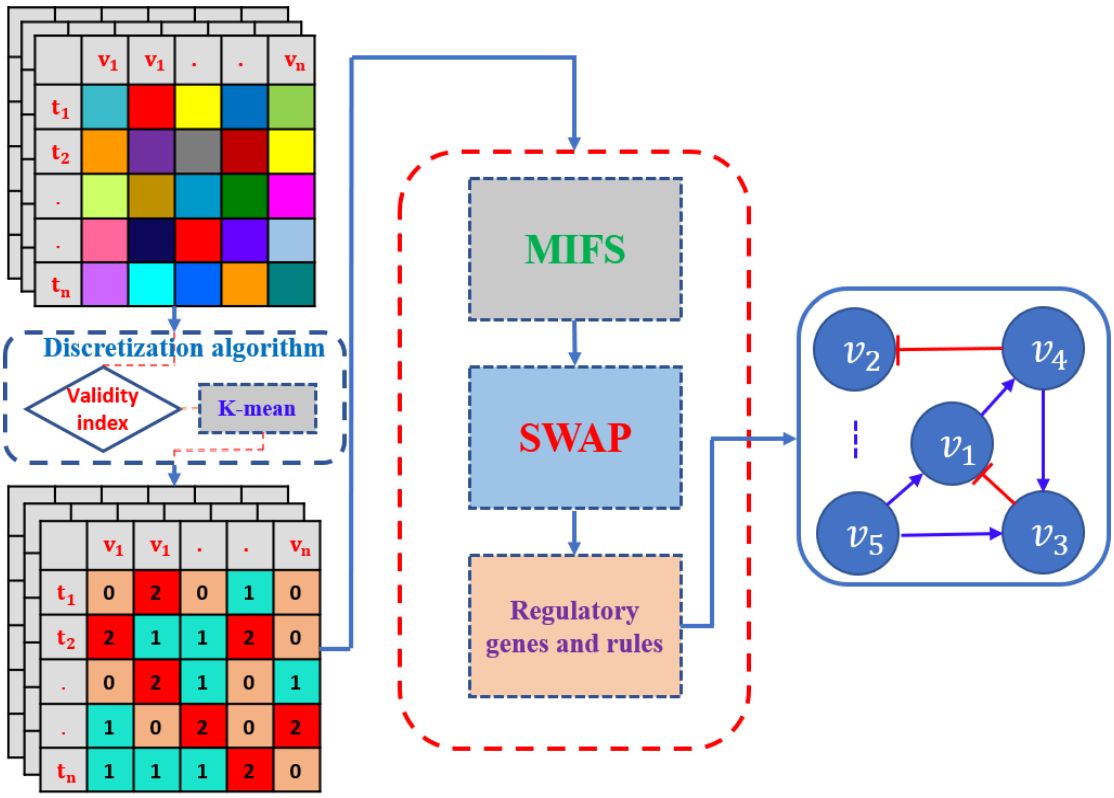


Figure 1. The framework of MIDNI algorithm. A real-valued time-series gene expression dataset is converted into a discretized expression dataset using the K-means discretization algorithm. K value is determined based on the validity index for each gene. The data is then fed into MIFS and SWAP subroutines. The result is the sets of regulatory genes for genes of the input network. Based on this result, we will reconstruct the network as a predictive network.

3.2.1. Discretization

Boolean discretization can speed up the inference process and some genes are suitable for such Boolean discretization. However, other genes are more appropriate for a higher-level discretization as shown in figure 2. In fact, there have been some previous studies to show the usefulness of level three discretization of gene expression. For example, Deep Neural Network [15] uses 1, 0 and -1 representing the genes value while calculating the individual gene's contribution. Some other methods [1-3]also applied a higher order discretization for genes whose values are denoted by 1, 0, and -1 meaning `upregulated', `no-change' and `downregulated'. Hence, the MIDNI employed a hybrid approach for data discretization with both levels of two

and three. To determine the optimal level of discretization, we used the validity index (more detail see literatures [18] [16, 17]), which is defined as follows:

$$Validity = \frac{Intra}{Inter},$$

where

$Intra = \frac{1}{P}\sum_{i=1}^{k}\sum_{x\in C_i}\|x - z_i\|^2$, and

$$Inter = \min_{i,j}(\|z_i - z_j\|)^2$$

where $i \neq j \in \{1, 2, \dots, k\}$, P is the number of observed timepoints in the gene, k is the number of the clusters, and z_i denotes the center of the cluster C_i. The optimal number of clusters will be chosen as to minimize the validity index.

More specifically, a gene is classified into two classes and three classes by using K-means clustering. Next, the validity index is calculated for each classification case. Final, the gene will be classified in a way that has a smaller corresponding validity index. This process is executed for all the genes of the network.

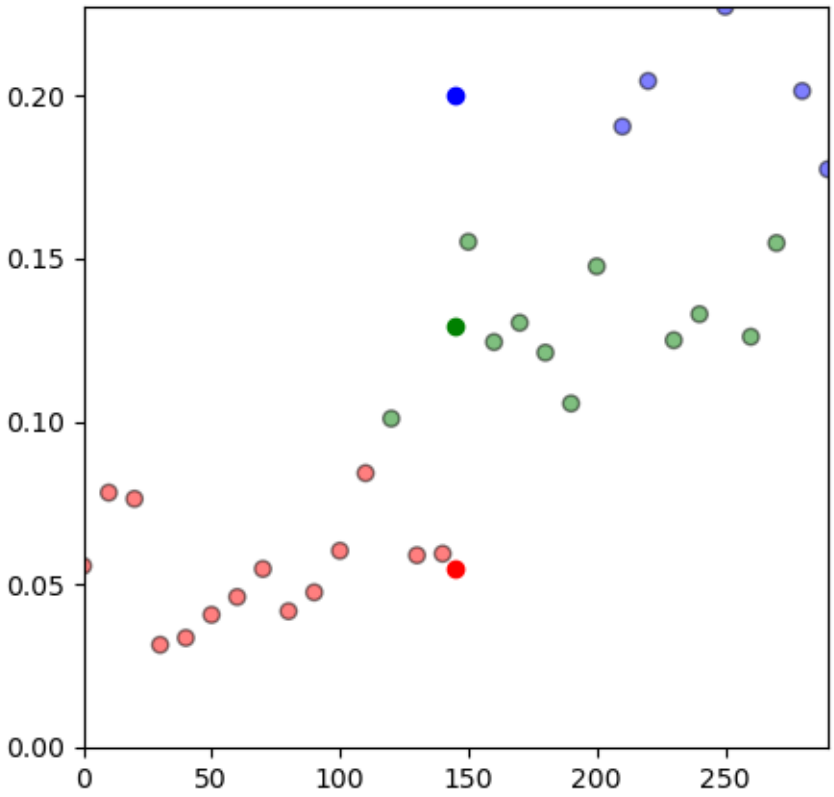
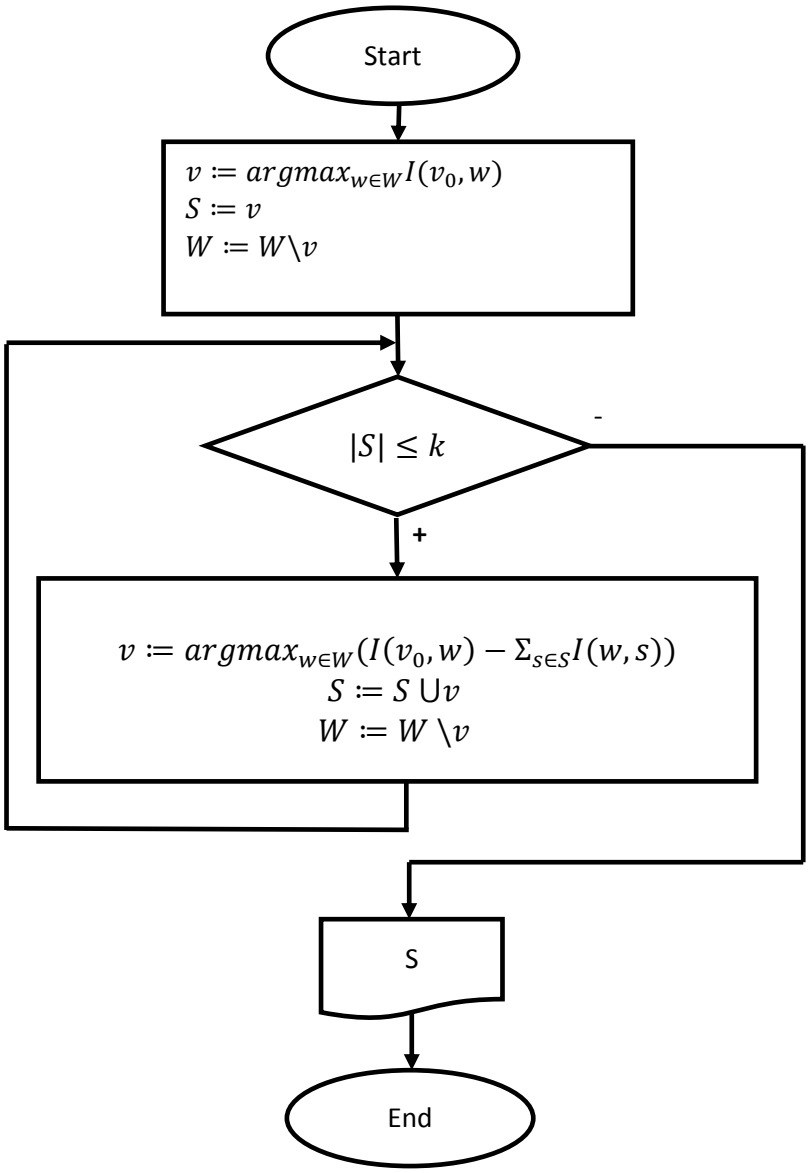


Figure 2. The real-valued distribution of genes that are more appropriate for a higher level of discretization. The x-axis shows the time points while y-axis denotes the real value of genes.

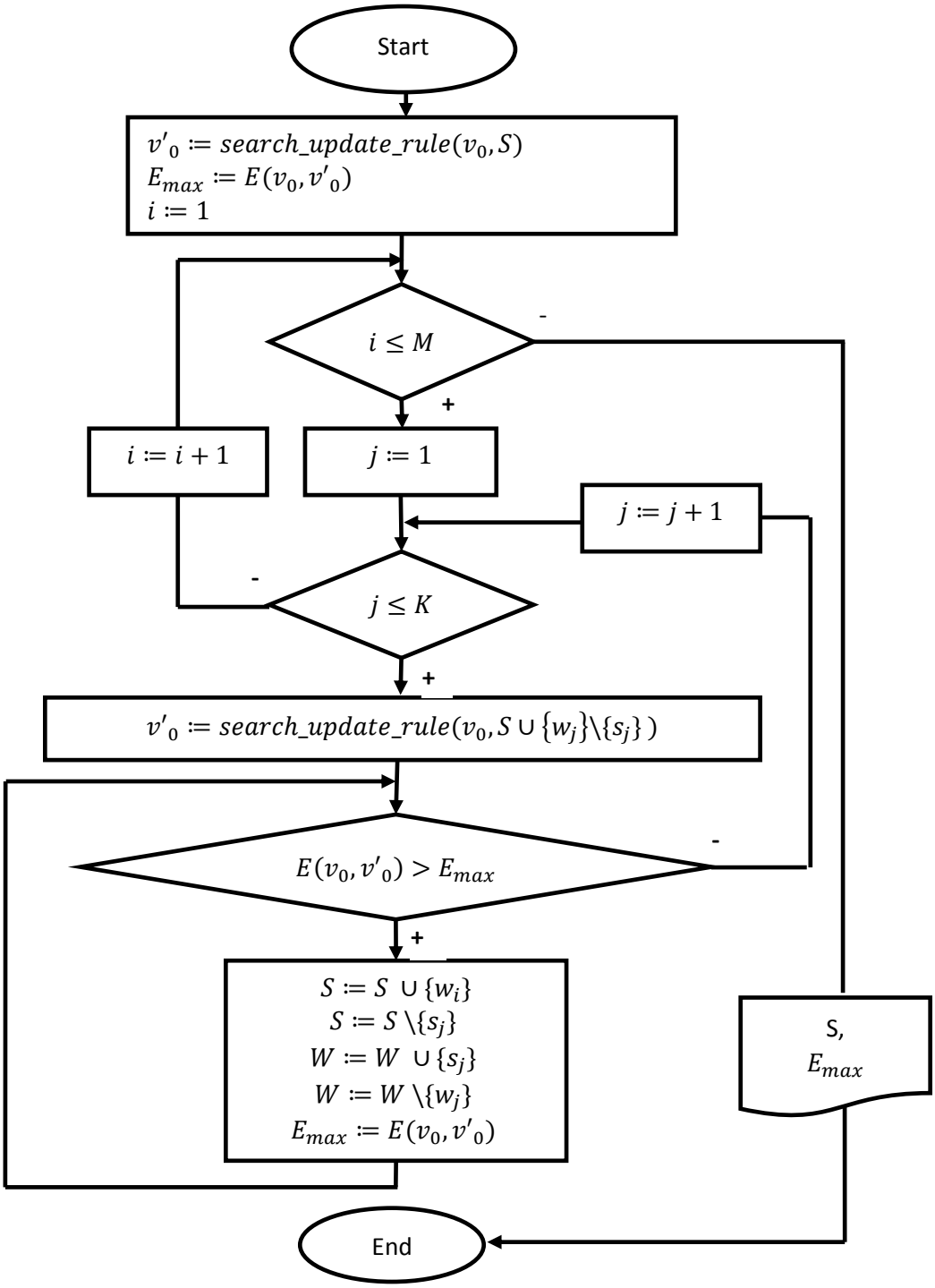
3.2.2. MIFS and SWAP subroutines

Given a discretized time-series gene expression dataset, MIDNI executes two subroutines, MIFS, and SWAP, in sequence. We note that they were modified from the similar subroutines in MIBNI (18) because the latter can handle only binary expression values. For each target gene v of the network, the MIFS subroutine searches an initial set of regulatory genes $S_v \subseteq V$ by evaluating an approximated multivariate mutual information (see Algorithm 1). Subsequently, the SWAP subroutine (see Algorithm 2) is used to iteratively search an improved set of regulatory genes starting at S_v . The *search_update_rule* function is a function that finds a set of update rules in the update table by which the input set S_v can perfectly transfer to the target gene v. More details about the *search_update_rule* function, see

an example in the Appendix. More specifically, the SWAP repeatedly tests a swapping between a selected candidate regulatory gene S_v (S in Algorithm 2) and the set of unselected genes $V \setminus S_v$ (W in Algorithm 2). The swapping process is repeated for every pair of variables in the selected and unselected sets of variables in the decreasing order of the mutual information with the target gene. As a result, the MIDNI returns a best-found set of regulatory genes for each target gene.



Algorithm 1. ***Subroutine*** $MIFS(v_0, W, k)$ *where* v_0 *- the target variable,* $W = \{w_1, w_2, \ldots, w_M\}$ *– the set of variables,* k *– the desired number of the input variables.*



Algorithm 2. ***Subroutine*** $SWAP(v_0, S, W)$ *where* v_0 *– the target variable;* $S = \{s_1, s_2, \dots, s_k\}$ *– the set of selected variables such that* $I(v_o, s_i) \geq I(v_0, s_j)$ *if* $i < j$ *for all* $s_i, s_j \in S$ *;* $W = \{w_1, w_2, \dots w_M\}$ *– the set of unselected variables such that* $I(v_0, w) \geq I(v_0, w_j)$ *if* $i < j$ *for all* $w_i, w_j \in W$.

.

Chapter 4

Experiments, Results, and Discussions

4.1. Experiments

The whole process can be divided into three stages: The first stage is to discretize the input data using the K-means algorithm. Then discretized data is fed into MIDNI, the result of this stage is the structure prediction of the network. This structure is compared with the real structure using a tool to evaluate the structure and dynamics accuracies. All stages are implemented with Python language and its libraries.

To validate the performance of MIDNI, we tested it with two time-series gene expression datasets: the artificial discretized and the Hill function-based real-valued ones. These two datasets are generated by also using Python language.

4.2. Results

4.2.1. Case study 1: Artificial discretized dataset

We generated an artificial dataset with 20 discretized network groups of |V| = 10, 20, ...,190, and 200. Each group includes 20 different random networks, and a total of 400 networks were tested in this simulation case. For each target gene, we randomly choose *k* genes from a set of all genes except the target while generating the structure of networks. The value *k* is also randomly generated ranging from 1 to the maximum number of regulators for each target gene (See Table 1). In addition, the state of every gene (node) was randomly initialized to a value of 0, 1, or 2 (i.e., three-level discrete value). It was updated over 29 time-steps by a discrete update function selected uniformly at random among a set of all possible update functions. Given a target gene

with k regulatory genes, there are 3^{3^k} possible update functions. In this way, we created the artificial discretized gene expression dataset. We also note that no discretization method is required for this dataset. (See Table 1 for setup). The inference result of MIDNI is shown in Figure 3. As shown in the figure, the dynamics accuracy was almost 1.0 for all tested networks. This perfect performance might be due to the maximum number of regulatory genes being limited to 8. With respect to the structural performance, the precision and recall decreased whereas the structural accuracy increased as the network size increased. This is because the numbers of candidate genes and true negative cases increase along with the increase of network size. We further examined the effect of the number of incoming links (*D*) on the performance (Figure 4). In a gold-standard network, the number of incoming links means the number of regulatory genes incoming to the. target gene and it eventually represents the difficulty of the inference problem. As shown in the figure, the number of incoming links ranged from 1 to 8 in the random gold-standard networks. When *D* was 1 or 2, our method showed almost perfect performance in terms of all dynamics and structural metrics. The structural performances continuously decreased as *D* gets larger than 2. However, the dynamics accuracy kept almost 1.0 even when $D = 8$. This implies that the inference problem is multimodal where a lot of optimal solutions exist in terms of the dynamic's accuracy. Taken together, our method stably inferred the near-optimal network which can show the dynamics similar to the gold-standard network, but the structural inference cannot be excellent because of the intrinsic multi-modality of the problem.

Parameters	Discretized dataset	Real-valued dataset
Size of networks (N)	10, 20, ..., 190, 200	50, 100, 200, 300
Noise rate	None	0, 5, 10
Time lag	1	1
Number of observed time points	30	30
Number of regulators in the network	$0.4 \times N$	$0.4 \times N$
Maximum regulators for each gene	4-8	4-8
Max expression (α_{max})	NA	0.1 - 1
Basal expression (β_0)	NA	0.1 - 1
Apparent dissociation constant (K_d)	NA	1 - 5
Hill coefficient (n)	NA	1 - 5

Table 1: The setting while generating datasets

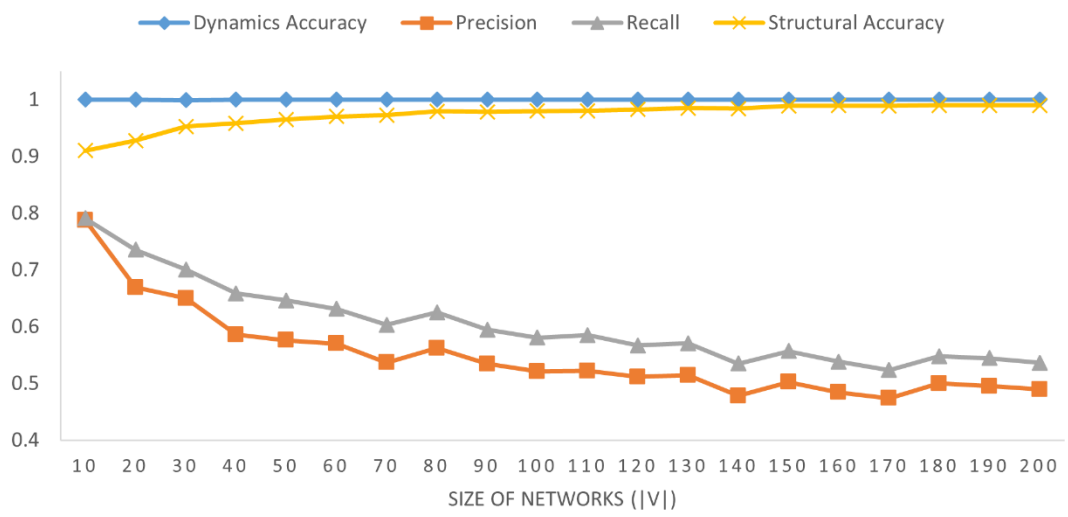


Figure 3. The performance of MIDNI on the artificial discretized dataset. The x-axis means the network size denoted by the number of genes in the network. For each size, the y-axis shows the average value of each performance metric over 20 random networks.

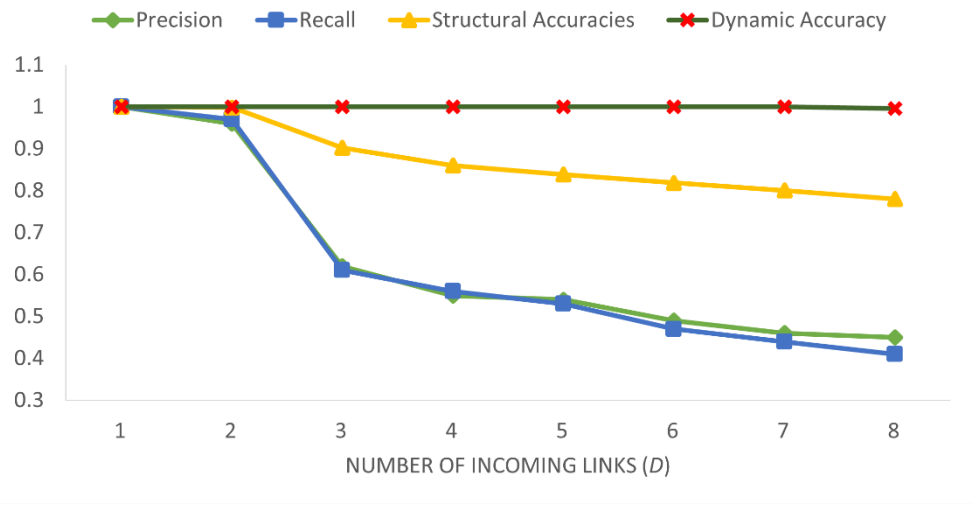


Figure 4. Inference performance according to the number of incoming links (D) in the gold-standard network.

4.2.2. Case study 2: Artificial real-valued dataset

MIDNI was tested with a real-valued time-series gene expression dataset generated by using the Hill function [19]. The dataset consists of four network groups of |V | = 50, 100, 200, and 300. Each group includes 20 different random networks, and thus a total of 80 networks were tested in this case study. The network structure was generated in the same way as for the artificial discretized dataset in the section 4.2.1. The used Hill function is described as follows (see Table 1 for the parameter setup):

$$[Protein] = \begin{cases} \alpha_{max}\left(\beta_0 + (1-\beta_0)\dfrac{[TF]^n}{K_d + [TF]^n}\right), for\ activation \\ \alpha_{max}\left(\beta_0 + (1-\beta_0)\dfrac{K_d}{K_d + [TF]^n}\right), for\ inhibition \end{cases}$$

where $\alpha_{max}, \beta_0, TF, K_d, n$ denote max expression value, basal expression, transcription factors, apparent dissociation constant and the Hill coefficient, respectively. As explained in section 3.2.1, a set of expression values of each gene was clustered by K-mean clustering where k equals to two or three depending on the distribution of expression values.

We first examined the proportion of genes which are three-level discretized, as shown in Figure 5. It noticed that apart from the 10-sized networks, the proportion of the two-level discretized genes and those of three-level discretized are relatively equal. This implies that ternary discretization is as frequent as binary one in our method.

We compared the inference performance of MIDNI with DBN, MICRAT, MIBNI and GENIE3 methods on the generated datasets. DBN [20] selects potential regulators by determining the time points of the initial changes in the expression (up- or down-regulation) of genes. MICRAT [21] builds an undirected graph representing the associations between genes based on the maximal information coefficient. Then, the directions of the edge in the undirected graph are determined by using average entropy. GENIE3 [22]is an approach inferring regulatory networks from expression data using tree-based methods and is the best performer in the DREAM4 In Silico Multifactorial challenge. MIBNI is our previous study that used Boolean values to infer the network.

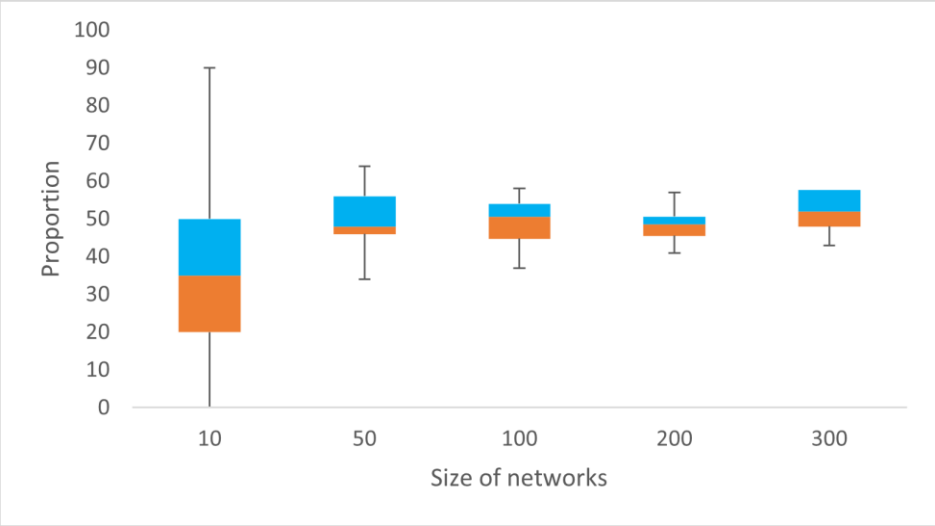


Figure 5. The proportion of three-level discretized genes in the networks.

We compared the performance of MIDNI with DBN, MICRAT, MIBNI and GENIE3 methods on the artificial Time-series gene expression datasets. DBN [20] selects

potential regulators by determining the time points of the initial changes in the expression (up- or down-regulation) of genes. MICRAT [21] builds an undirected graph representing the associations between genes based on the maximal information coefficient. Then, the directions of the edge in the undirected graph are determined by using average entropy. GENIE3 [22] is an approach inferring regulatory networks from expression data using tree-based methods and is the best performer in the DREAM4 In Silico Multifactorial challenge. MIBNI is our previous study that used Boolean values to infer the network.

4.2.3. Structural accuracy analysis

We classified all of the target genes into eight groups according to the number of incoming links (D), which represents the number of regulatory genes for each target gene ranging from 1 to 8 as shown in Figure 6. As shown in the figure, the structural accuracy values of all methods reduced as D increased. This is because the number of incoming links represents the degree of difficulty of the inference problem. In the case of D = 1, all methods almost always found the optimal solution. However, as D increases, their performances linearly reduced. We note that our method showed the best structural accuracy in all cases except for D = 1 and 2 cases in the smallest network (|V | = 50). In particular, the performance gap was clearer when the network size and D are relatively large. This result can explain the efficiency of our method in a large-scale and difficult-to-solve inference problem. We also note that multi-level discretization was useful by comparison with our previous method, MIBNI.

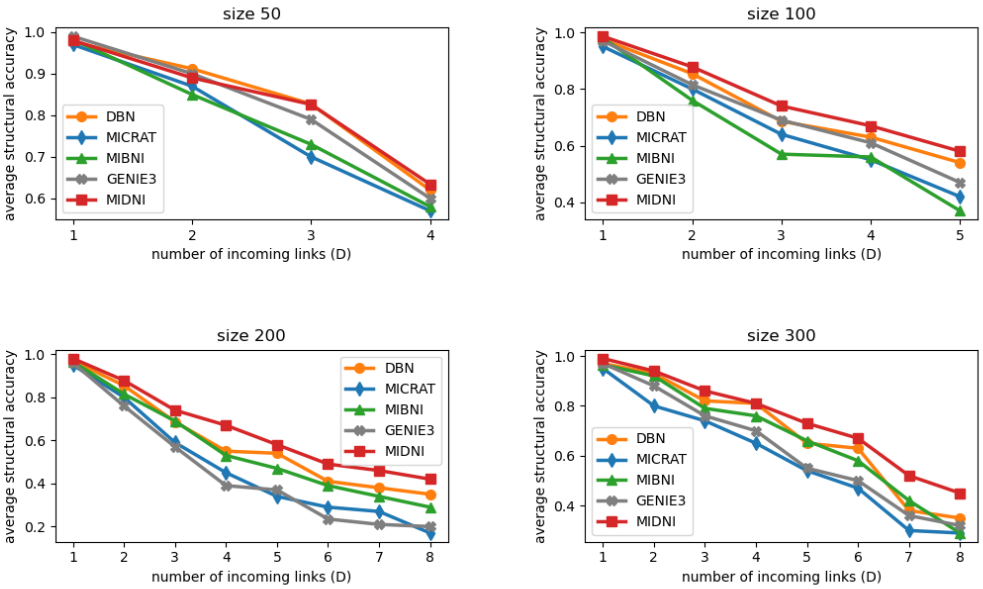


Figure 6. Structural performance on the random real-valued networks. Four size groups of networks (|V | = 50, 100, 200, and 300) were considered, and 20 random networks were generated for each group. All target genes are grouped according to the number of incoming links (D) in the gold-standard network, and the x-axis means D values. The y-axis value shows the average structural accuracy.

4.2.4. Dynamics accuracy analysis

We notice that the different structures of the networks may produce the same underlying dynamics. Hence, it is also essential to verify the network inference performance in terms of the dynamics accuracy. We examined the dynamics accuracies of the inferred networks by MIDNI, DBN, MICRAT, MIBNI, and GENIE3 over the networks. Similar to structural accuracy, we also analyze the influence of the regulatory genes number of each target gene on dynamic accuracy for each method. The experiments are shown in Figure 7. As shown in the figure, most methods showed almost 1.0 dynamics accuracy when D = 1. As D increases, the inference problems get more difficult, and thus their dynamic accuracies decrease. However, MIDNI showed the best performance in all cases except for D = 3 in the network group of |V | = 300. Specifically, in the case of D = 8 for the network group

of sizes 200 and 300, the dynamics accuracy of MIDNI was about 0.6 whereas those of the other methods were around 0.5. This result also proves that our method has a notable inference ability for the more complex and larger networks.

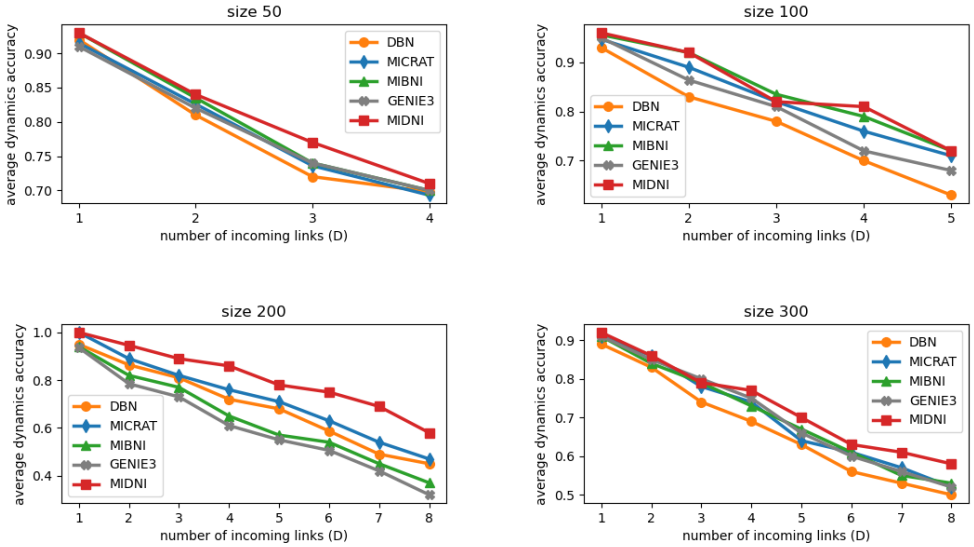


Figure 7. Dynamical performance on the random real-valued networks. Four size groups of networks (|V | = 50, 100, 200, and 300) were considered, and 20 random networks were generated for each group. All target genes are grouped according to the number of incoming links (D) in the gold-standard network, and the x-axis means D values. The y-axis value shows the average dynamics accuracy.

4.2.5. Running time

To compare the running time of MIDNI and other methods, the average running time was examined over a total of 80 random networks that were mentioned in the section 4.2.2. The experiments are conducted on a PC with AMD Ryzen 5 3400G 3.70 GHz CPU and 16 GB RAM. The result is shown in Figure 8. In the figure, the Y-axis value means the average running time in milliseconds along with the network group with four different network sizes. Although MIDNI was clearly slower than DBN and MICRAT it was comparable with GENIE3 and MIBNI. MIDNI is slightly faster than MIBNI. This is because we have made some optimizations of computational commands of the MIFS, and SWAP subroutines used in MIDNI from the original

one. For example, we used build-on libraries such as *sklearn* and *scipy* to calculate the mutual information score and the entropy metrics. Build-on functions are well-organized and more optimal than self-implemented ones. We also used computation on vectors which are executed much faster than those done by the *for* loop. Considering that the structural and dynamics inference accuracy of MIDNI outperformed the other methods, it is desirable to apply our method when the high accuracy is needed by sacrificing the running time cost.

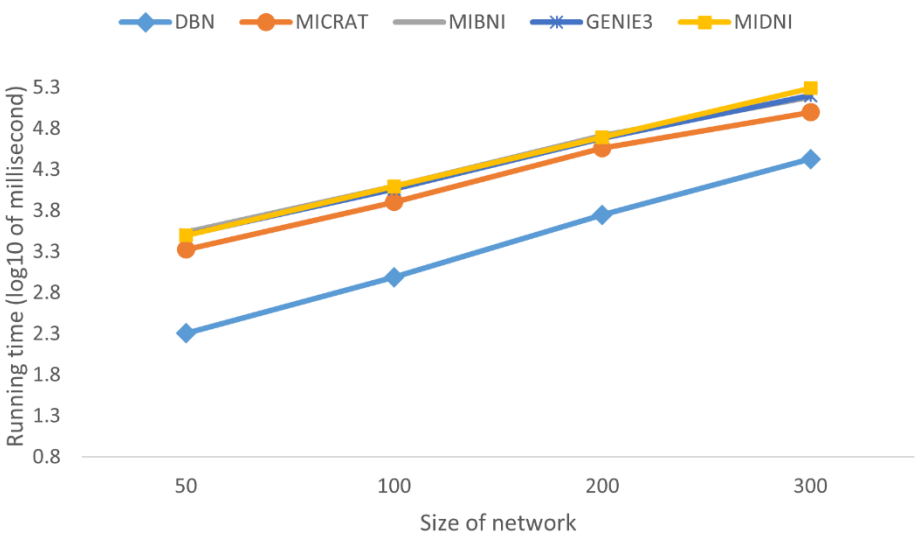


Figure 8. Comparison of running time.

4.3. Discussions

In this study, we proposed a new network inference method from time-series gene expression data, called MIDNI. It first converts the real-valued expression into multiple-level discretized expression adequately according to the distribution of expression values. We note that most previous discretization-based inference methods have employed a Boolean network model, and this can limit the inference performance due to the expression simplicity. In this study, our method outperformed the compared methods in both terms of structural and dynamical accuracies,

particularly in more dense and larger networks. Despite a successful result, MIDNI has some limitations to be improved. First, the precision and the recall values of MIDNI are not so high. We infer the reason because MIDNI relies on the mutual information to select a set of candidate regulatory genes which directly interact with a target gene. However, some genes may interact with others via an intermediate gene, and this can lead to erroneous selections. This implies that a hybrid approach to combine our method with a priori

knowledge from the biological database can make more accurate predictions. Second, MIDNI uses the correlation coefficient to determine the type and direction of an interaction, which can also cause a reduction of inference accuracy. Finally, MIFS and SWAP subroutines used in MIDNI are greedy algorithms, so the performance can be improved if more efficient search algorithms are combined.

Chapter 5

Conclusion & Future Studies

5.1. Conclusion

This research tried to solve one of the most challenging problem in the biological field that is regulatory gene inference. The distinguishes of the proposed method and the others is that instead of using the raw-data (real-values) or the Boolean, a multiple level data representation approach is used. As a result, the method not only avoids information loss during data representing, moreover, speeds up the inference process. It also gives better results than the previous methods both in terms of structural and dynamics accuracy, especially for the large size of networks. Here are some of the problems that have been solved by our method:

- Determine the optimal number of clusters (discretization level) when conducting data representation for each gene in the network. This number depends on the distribution of the gene.

- The algorithm finds the potential candidate regulatory gene by applying not only the mutual information between two genes but also the approximate dependency of the candidate gene to the set of the prior selected candidates. This made our algorithm different from major other approaches that consider only the pair-gene information.

- MIDNI does not only improve the structural accuracy between the input and predictive networks, but also was implemented to increase the dynamics consistency within genes in the network.

5.2. Future Studies

Although the proposed method outperformed other compared methods, it still has some limitations as mentioned in 4.2 section which should be done in the next stage of this research. In the future, I intend to carry out some work as follows:

- Develop a new subroutine that can adapt to multimodal problem where there are a lot of optimal solutions.

- Apply other metric to determine the interaction direction of genes during the inference instead of correlation coefficient which is relatively simple and inadequate for enhancing the accuracy.

- Speed up the inference process by applying optimized search algorithms or parallel computation techniques.

- Conduct an experiment of data representation by using other algorithms to compare with K-mean.

- Try to infer the network with other kind of data like the steady state (in this study using only time-series) and combine with prior knowledge data to improve the inference accuracy.

Bibliography

1. S. Madeira, A.L.O., *An Evaluation of Discretization Methods for Non-Supervised Analysis of Time-Series Gene Expression Data.* Computer Science, 2005.
2. Gallo CA, C.J., Ponzoni I, *Biological Knowledge Discovery Handbook: Preprocessing, Mining, and Postprocessing of Biological Data*. Vol. 36. 2013, In: Elloumi M, Zomaya AY (eds).
3. Ji L, T.K., *Mining gene expression data for positive and negative co-regulated gene clusters.* Bioinformatics, 2004: p. 2711-18.
4. Zhang, B. and S. Horvath, *A general framework for weighted gene co-expression network analysis.* Stat Appl Genet Mol Biol, 2005. **4**: p. Article17.
5. Butte, A.J. and I.S. Kohane, *Mutual information relevance networks: functional genomic clustering using pairwise entropy measurements.* Pac Symp Biocomput, 2000: p. 418-29.
6. Faith, J.J., et al., *Large-scale mapping and validation of Escherichia coli transcriptional regulation from a compendium of expression profiles.* PLoS Biol, 2007. **5**(1): p. e8.
7. Zoppoli, P., S. Morganella, and M. Ceccarelli, *TimeDelay-ARACNE: Reverse engineering of gene networks from time-course data by an information theoretic approach.* BMC Bioinformatics, 2010. **11**: p. 154.
8. Meyer, P.E., et al., *Information-theoretic inference of large transcriptional regulatory networks.* EURASIP J Bioinform Syst Biol, 2007: p. 79879.
9. Harri Lähdesmäki, I.S., Olli Yli-Harja *On Learning Gene Regulatory Networks Under the Boolean Network Model.* Machine Learning, Springer, 2003. **52**: p. 147-167.
10. Haider, S. and R. Pal, *Boolean network inference from time series data incorporating prior biological knowledge.* BMC Genomics, 2012. **13 Suppl 6**: p. S9.
11. Barman, S. and Y.K. Kwon, *A novel mutual information-based Boolean network inference method from time-series gene expression data.* PLoS One, 2017. **12**(2): p. e0171097.
12. Barman, S. and Y.K. Kwon, *A Boolean network inference from time-series gene expression data using a genetic algorithm.* Bioinformatics, 2018. **34**(17): p. i927-i933.
13. Barman, S. and Y.K. Kwon, *A neuro-evolution approach to infer a Boolean network from time-series gene expressions.* Bioinformatics, 2020. **36**(Supplement_2): p. i762-i769.
14. MacQueen, J., *Some methods for classification and analysis of multivariate observations.* Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, 1967. **1**: p. 281-297.
15. TaeJin Ahn, T.G., Chan-hee Lee, SungMin Kim, Kyullhee Han, Sangick Park, Taesung Park, *Deep Learning-based Identification of Cancer or Normal Tissue using Gene Expression Data.* IEEE International Conference on Bioinformatics and Biomedicine (BIBM), 2019: p. 1748-1752.
16. S. Ray, R.H.T., *Determination of number of clusters in K-means clustering and application in colour image segmentation.* Proceedings of the 4th International Conference on Advances in Pattern Recognition and Digital Techniques (ICAPRDTÕ99), 1999: p. 27-29.

17. M. Halkidi, Y.B., M. Vazirgiannis, *Clustering algorithms and validity measures.* Proceedings of SSDBM Conference, Virginia, USA, 2001.
18. Judong Shen, S.I.C., E. Stanley Lee, Youping Deng, Susan J. Brown, *Determination of cluster number in clustering microarray data.* Applied Mathematics and Computation, 2005. **169**(2): p. 1172-1185.
19. Santillan, M., *On the Use of the Hill Functions in Mathematical Models of Gene Regulatory Networks.* Mathematical Modelling of Natural Phenomena, 2008. **3**(2): p. 85-97.
20. Zou, M. and S.D. Conzen, *A new dynamic Bayesian network (DBN) approach for identifying gene regulatory networks from time course microarray data.* Bioinformatics, 2005. **21**(1): p. 71-9.
21. Yang, B., et al., *MICRAT: a novel algorithm for inferring gene regulatory networks using time series gene expression data.* BMC Syst Biol, 2018. **12**(Suppl 7): p. 115.
22. Huynh-Thu, V.A., et al., *Inferring regulatory networks from expression data using tree-based methods*. PLoS One, 2010. **5**(9).

Appendix

Search_update_rule

Assume that we have the target gene v and an initial set of regulatory genes S_v that consists of two regulatory genes v_1 and v_2. The value of genes is shown in the table below:

Target gene v	Set of regulatory genes S_v	
	v_1	v_2
0	**1**	**1**
2	**0**	**1**
1	**1**	**1**
2	**1**	**2**
0	**0**	**2**
0	**0**	**2**
2	**1**	**0**

With two regulatory genes for the target gene, so we have the update table as follows:

input		output	input		output	input		output
0	**0**	**0**	**0**	**0**	**1**	**0**	**0**	**2**
0	**1**	**0**	**0**	**1**	**1**	**0**	**1**	**2**
1	**0**	**0**	**1**	**0**	**1**	**1**	**0**	**2**
1	**1**	**0**	**1**	**1**	**1**	**1**	**1**	**2**
0	**2**	**0**	**0**	**2**	**1**	**0**	**2**	**2**
2	**0**	**0**	**2**	**0**	**1**	**2**	**0**	**2**
2	**2**	**0**	**2**	**2**	**1**	**2**	**2**	**2**
1	**2**	**0**	**1**	**2**	**1**	**1**	**2**	**2**
2	**1**	**0**	**2**	**1**	**1**	**2**	**1**	**2**

With the time lag p = 1, the set of update rules as the output of the *search_update_rule* in the update table is shown in table below:

input		output
1	**1**	**2**
0	**1**	**1**
1	**2**	**0**

With the set of update rules above, we can transfer any values of v_1 and v_2 to v (except two last cases due to having different output with the same input) as shown by the color highlight in the table below:

Target gene v	Set of regulatory genes S_v	
	v_1	v_2
0	**1**	**1**
2	**0**	**1**
1	**1**	**1**
2	**1**	**2**
0	**0**	**2**
0	**0**	**2**
2	**1**	**0**