THE REPUBLIC OF TURKEY
BAHCESEHIR UNIVERSITY

C3NET ALGORITHM USING DYNAMIC BAYESIAN NETWORK

Master's Thesis

MOHAMMED ABDULGHANI TAHA

İSTANBUL, 2013

THE REPUBLIC OF TURKEY

BAHCESEHIR UNIVERSITY

THE GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCE COMPUTER ENGINEERING

C3NET ALGORITHM USING DYNAMIC BAYESIAN NETWORK

Master's Thesis

MOHAMMED ABDULGHANI TAHA

Supervisor: ASSIST. PROF. DR. GOKMEN ALTAY

İSTANBUL, 2013

ACKNOWLEDGEMENTS

I would like to express my gratitude to my supervisor Assoc. Prof. Dr. Gokmen ALTAY for the useful comments, remarks and engagement through the learning process of this master thesis. Further more I would like to thank my parents, brothers and sisters.

ABSTRACT

C3NET ALGORITHM USING DYNAMIC BAYESIAN NETWORK

Mohammed Abdulghani Taha
M.S. Department of Computer Engineering
Supervisor: Assist. Prof. Dr. Gokmen ALTAY

April 2013, 33 pages

Finding causal interactions between genes is one of the most important topics in bioinformatics. Many gene regulatory network inference (GRNI) algorithm has been introduced for this aim. In this study, we use C3NET algorithm and G1DBN algorithm.

C3NET algorithm's inferred gene network is undirected. G1DBN algorithm's inferred gene network is directed but it's too slow when applied to large expression data, it takes too much time to infer directed gene networks.

Our approach solves both direction and time by applying Dynamic Bayesian Network to the inferred gene network of C3NET to make the inferred gene network directed. So our approach composed of two steps, in the first step decreases the interaction probability of genes by C3NET algorithm, in the second step applies Dynamic Bayesian network to each pair interaction of genes and make the undirected edges to directed edges.

Keywords: Dynamic Bayesian Network, Directed Acyclic Graph, networks inference, conditional independence, time series modeling.

.

ÖZET

DİNAMİK BAYES AĞLARI KULLANARAK C3NET ALGORİTMASI

Mohammed Abdulghani Taha

Yüksek Lisans, Bilgisayar Mühendisliği Bölümü

Tez Yöneticisi: Doç. Dr. Gökmen ALTAY

Nisan 2013, 33 sayfa

Genler arasındaki nedensel ilişkileri bulma biyoinformatik'te en önemli konulardan biridir. Birçok gen düzenleyici ağ çıkarım (GRNI) algoritmasları bu amaçla gelişitirilmiştir. Bu çalışmada, C3NET algoritma ve G1DBN algoritması kullanırılıyor.

C3NET algoritmanın anlaşılmaktadır gen ağı yönsüzdür. G1DBN algoritmanın anlaşılmaktadır gen ağı yönlüdür ama büyük veriler'de uygulandığında çok yavaş çalışıyor, yönlendirilmiş gen ağları bulması için çok fazla zaman gerektirir.

Yaklaşımımız anlaşılmaktadır gen ağı yapmak için C3NET ve Dinamik Bayes Ağı uygulayarakö yön ve zaman gecikmesini çözüyor. Bizim yaklaşım iki adımdan oluşuyor, ilk adımda C3NET algoritması tarafından genlerin etkileşimi olasılığı azalır, İkinci aşamada genlerin her çift etkileşimi Dinamik Bayes ağ geçerlidir ve yönsüz ağı yönlü ağa çevirir.

Anahtar Kelimeler: Dinamik Bayes Ağ, Yönetmen Mercury Graph, çıkarım ağları, koşullu bağımsızlık, zaman serisi modelleme.

CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVATIONS

BN :	Bayesian Network
DBN :	Dynamic Bayesian Network
DAG:	Directed Acyclic Graph
GRNI:	Gene Regulatory Network Inference
DPI:	Data Processing Inequality
OLAP:	On-line Analytical Processing
GGMs :	Graphical Gaussian Models
MI:	Mutual Information

1. INTRODUCTION

1.1 PROBLEM DEFINITION

The structure and working mechanisms of molecules in cells of an organism will be clear while, a logical understanding of biological and biomedical problems is performed. The interaction among genes and gene product displays the gene networks of an individual, e.g., the transcriptional regulatory network, protein network or metabolic network (Lebre, 2012) (Lebre, 2009). The blueprints of dynamical processes within cells are represented by these networks (Altay & Emmert-Streib, 2011). Because of this, the inference of gene networks from experimental data is called as one of the most important targets of the post-genomic era and in system biology (Altay & Emmert-Streib, 2010).

An accurate detection of molecular interaction is allowed by classical molecular biology approaches (Altay & Emmert-Streib, 2010). In the early 1940 s BEADLE and TATUM (Emmert-Streib & Dehmer, 2010) focused on the assumption of the one gene-one enzyme hypothesis which caused the study of the molecular biology for decades, but the current focus is on the systems properties of interacting genes (Emmert-Streib, 2011) (Vidal, 2009). Since the high-throughput data has been appeared, the study of the behavior of such systems is focused (Altay & Emmert-Streib, 2011). For example, a wealth information about the expression of genes are provided by microarray experiments, these information can be utilized by statistical analysis methods in order to investigate data systematically (Dudoit, Shaffer, & Boldrick, 2003) (Speed, 2003). One of the important thing that increased the interest of the microarray analysis is the causal interactions among thounsands of genes (Li & Gui, 2006) (Xing & van der Laan, 2005). Here by causal, the direct interactions among genes that correspond to experimentally verifiable biochemical interactions is meant (Altay & Emmert-Streib, 2011).We mean, relationship between two gene is searched, e.g., “gen i activates gene j”. It is known that most of the genes whose expression has been monitored using microarrays are not present in the temporal evolution of the system (Lebre, 2009). So the determination of the few ‘active’ genes and the relationships between them is required. In summary, we want to estimate a network that contains the dependence relationships.

1.2 LITERATURE REVIEW

To infer these type of networks static modeling first was described which are not oriented network. The relevance network (Butte, Tamayo, Slonim, Golub, & Kohane, 2000) or correlation network (Steuer, Kurths, Fiehn, & Wechwerth, 2003) was one of the first tools used to infer interactions between genes (Lebre, 2009). This method calculates pair wise mutual information values among all genes and deletes the edges among genes which have mutual information values that are not statistically significant (Altay & Emmert-Streib, 2010). Also it is known as the covariance graph (Cox & Wermuth, 1996) in graphical models theory, this undirected graph shows the pair-wise correlation between genes. There is an undirected edge between two nodes (variables) whenever there is a correlation, this topology is taken from the covariance matrix between gene expression levels (Lebre, 2009). However, the relation between two nodes could be caused by linkage with other variables. This generates fake edges due to indirect dependence relationships (Lebre, 2009).

As a result, there has been interest in the concentration graph (Lauritzen S. L., 1996), also mentioned the covariance selection model, which manipulates the conditional dependence structure between gene expression using Graphical Gaussian Models (GGMs). Let $Y = (Y^i)_{1\leq i\leq p}$ be a Gaussian vector representing the expression levels of p genes (Lebre, 2009). Since they are conditionally dependent, an undirected edge is drawn between two variables Y^i and Y^j (Lebre, 2009). The theory GGMs can be used only when the number of measurements n is much higher than the number of variables p (Lebre, 2009). However most of the microarray gene expression datasets are opposite where the number of variables p is much higher than n. Thus, the interest in “small n, large p” forced the development of more alternatives (Schafer & Strimmer, 2005) (Schafer & Strimmer, 2005) (Waddell & Kishino, 2000) (Waddell & Kishino, 2000) (Toh & Horimoto, 2002) (Toh & Horimoto, 2002) (Wu, Ye, & Subramanian, 2003) (Wang, Myklebost, & Hovig, 2003).

Gene regulatory network inference (GRNI) algorithms are an essential means to gather genome-scale causal interaction networks (Emmert-Streib, 2011). More of GRNI methods are information theory based approaches (Butte & Kohane, 2000) (Watkinson, Liang, Wang, Zheng, & Anastassiou, 2009). More of the such approaches are inference methods which are based on calculation of mutual information (MI) values (Butte & Kohane, 2000) (Kraskov, Stagbaur, & Grassberger, 2004) (Margolin, et al., 2006). Unlike Pearson correlation coefficient, MI value can detects linear and non-linear effects among gene pairs, so this is more comfortable in a genome context (Li W. , 1990) (Steuer, Kurths, Daub, Weise, & Selbig, 2002).

So many methods are appeared. Another (GRNI) is ARACNE (Algorthim for the Reconstruction of Accurate Cellular Networks) (Margolin, et al., 2006) which is similar to RN. In ARACNE, the data processing inequality (DPI) (Cover & Thomas, 1991) is used to eliminate the least significant edge of a triplet of genes, which is equal to the lowest mutual information value thereof (Altay & Emmert-Streib, 2010). Since ARACNE can contain at most as many interactions as inferred by RN, gives a better estimation of the inferred network (Altay & Emmert-Streib, 2010). CLR (Context Likelihood of Relatedness) is another method similar to RN (Faith, et al., 2007) which has a sensitive estimator for the connection among genes, this is done by converting mutual information estimates into z-score like values. The final GRNI method we illustrate is MRNET (maximum relevance/minimum redundancy Network) (Meyer, Kontos, & Bonternpi, 2007). This method provide the maximum relevance/minimum redundancy (MRMR) feature selection method (Ding & Peng, 2005) (Tourassi, Frederick, Markey, & Floyd, 2001). A new GRNI algorithm, C3NET (Altay & Emmert-Streib, 2010), is developed. C3NET is also based on MI, and has been compared with other GRNI algorithms (Altay & Emmert-Streib, 2010). We illustrate it in section Data and Methods. Because our is related with C3NET. All the method we illustrate do not have an accurate description of the interactions. For e.g, there is no direction between genes. Unlike the other algorithms, Bayesian networks (BNs) model (Friedman, Linial, Nachman, & Pe'er, 2000) is directed relationships.

BN model is introduced by a Directed Acyclic Graph (DAG) and the set of conditional probability distributions of each variable given its parents in the DAG (Pearl, 1988) based on a probabilistic measure (Lebre, 2009). Static BNs has an careful restriction that gives the structure of genetic networks, this restriction is because of its acyclicity constraint (Lebre, 2009). This limitation can be solved by providing Dynamic Bayesian network (DBNs) which is used for analyzing gene expression time series by Friedman et al.. However, the microarray gene expression datasets are very huge, it takes long time to estimate the causal interaction between genes. So here our approach solve the weakness by combining GRNI method C3NET (Altay & Emmert-Streib, 2010) which decrease the number of genes and applying DBN method to them. In our approach we use two packages, one is for C3NET algorithm which is c3net (Altay & Emmert-Streib, 2011) package, second is for DBN which is G1DBN (Lebre, 2012).

2. DATA & METHODS

2.1 BAYESIAN NETWORK

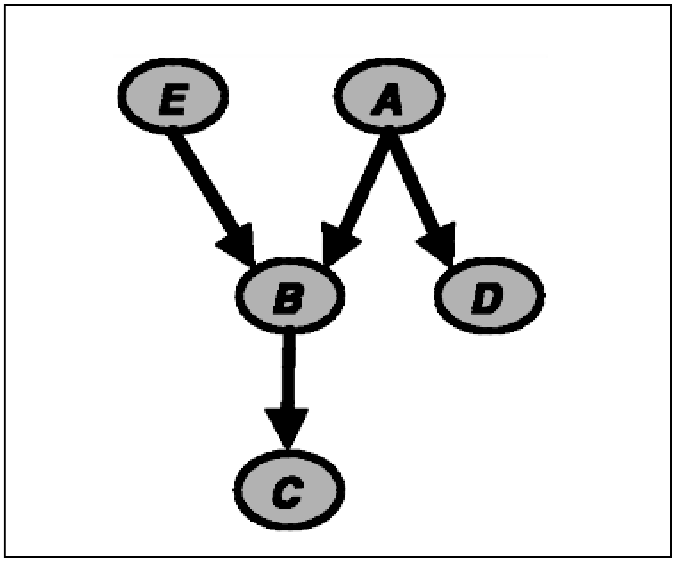
Bayesian networks are very important in many areas of biological sciences like in cellular networks (Friedman, 2004), modeling protein signaling pathways (Sachs, Perez, Pe'er, Lauffenburger, & Nolan, 2005), systems biology, data integration (Sachs, Perez, Pe'er, Lauffenburger, & Nolan, 2005), classification (Bradford, Needham, Bulpitt, & Westhead, 2006), and genetic data analysis (Beaumont & Rannala, 2004). Bayesian networks are suitable for combining domain knowledge and data, expressing causal relationships and learning incomplete datasets by using probability theory (Needham, Bradford, Bulpitt, & Westhead, 2007).

Bayesian networks have been used in many areas, e.g; they have been used in On-line Analytical Processing (OLAP) performance enhancement (Scutari, Learning Bayesian Networks with the bnlearn R Package, 2010), medical service performance analysis (Scutari, 2010) (Acid, de Campos, Fenandes-Luna, Rodriguez, & Salcedo, 2004), gene expression analysis (Friedman, Linial, Nachman, & Pe'er, 2000), breast cancer prognosis and epidemiology (Holmes & Jain, 2008). Essential tool for analyzing gene expression are Bayesian networks.

2.1.1Distributions With Bayesian Networks

A finite set $X = \{X_1, \dots X_n\}$ of random variables are considered, where each variable X_i may have value x_i from domain $Val(X_i)$ (Friedman, Linial, Nachman, & Pe'er, 2000). Capital letters like X, Y, X, are used for variable names and lowercase letters are used like x, y, z, to illustrate values taken by those variables (Friedman, Linial, Nachman, & Pe'er, 2000). Boldface capital letters X, Y, X are used for sets of variables, and boldface lowercase letters x, y, z, are used for the assignments of values to the variables in these sets (Friedman, Linial, Nachman, & Pe'er, 2000). $I(X; Y|Z)$ is marked to mean X is independent of Y conditioned of Z: $P(X|Y, Z) = P(X|Z)$ (Friedman, Linial, Nachman, & Pe'er, 2000).

Figure 2.1 A simple Bayesian network structure.The Conditional independence statements $\boldsymbol{I(A;E)}$, $\boldsymbol{I(B;D \mid A,E)}$, $\boldsymbol{I(C;A,D,E, \mid B)}$, $\boldsymbol{I(D;B,C,E \mid A)}$, $\boldsymbol{and\, I(E;A,D)}$. The Joint distribution: $\boldsymbol{P(A,B,C,D,E) = P(A)P(B|A,E)P(C|B)P(D|A)P(E)}$.



Source: Friedman, N., Linial, M., Nachman, I., & Pe'er, D. (2000). *Using Bayesian networks to analyse expression data.*

A Bayesian network is a joint probability distribution representation (Friedman, Linial, Nachman, & Pe'er, 2000). The representation is composed of two components (Friedman, Linial, Nachman, & Pe'er, 2000). G is the first component which represents a directed acyclic graph (DAG) where its vertices are random variables $X_1 \ldots, X_n$ (Friedman, Linial, Nachman, & Pe'er, 2000). θ is the second component which defines the conditional distribution for each variable, where its parent are given in G (Friedman, Linial, Nachman, & Pe'er, 2000). A unique distribution on $X_1 \ldots, X_n$ is specified by these two components (Friedman, Linial, Nachman, & Pe'er, 2000). Conditional independence assumptions that allow the joint distribution to be decomposed is represented by the graph G. The graph G simulates the Markov Assumption: Xi variables are independent and have a parent in G (Friedman, Linial, Nachman, & Pe'er, 2000). Properties of conditional independencies and chain rule of probabilities is applied for any joint distribution that satisfies markov assumption which represented by the product form:

$$P(X_1,\ldots,X_n) = \prod_{i=1}^{n} P\left(X_i \middle| Pa^G(X_i)\right), \qquad (1)$$

Where $Pa^G(X_i)$ is the set of parents of X_i in G (Friedman, Linial, Nachman, & Pe'er, 2000). Figure 1 illustrate a simple example of a graph G and the lists of the Markov independencies (Friedman, Linial, Nachman, & Pe'er, 2000).

As in (1), a graph *G* provides a product form. To specify the fully joint distribution, the conditional distributions in the product form is needed to be specified (Friedman, Linial, Nachman, & Pe'er, 2000). This is will be the second part of the Bayesian network, which describes these conditional distributions, $P(X_i|Pa^G(X_i))$ for each variable X_i (Friedman, Linial, Nachman, & Pe'er, 2000). These distributions will be denoted by the parameter θ (Friedman, Linial, Nachman, & Pe'er, 2000).

Conditional distribution is represented according to the variable types:

a) **Discrete variables.** $P(X \mid U_1,\ldots,U_k)$ can be represented as a table provides the probability of values for *X* for each joint assignment to $U_1,\ldots,U_k$, while the values of *X* and $U_1,\ldots,U_k$ are discrete (Friedman, Linial, Nachman, & Pe'er, 2000).

b) **Continuous variables.** Since the variables of *X* and $U_1,\ldots,U_k$ real valued, all possible densities can not be represented (Friedman, Linial, Nachman, & Pe'er, 2000). Gaussian distribution is used for multivariate continuos distributions (Friedman, Linial, Nachman, & Pe'er, 2000). So here the conditional density of *X* with its parents represented as follow:

$$P\left(X\middle|u_{1,\ldots,}u_k\right) \sim N\left(a_0 + \sum_i a_i \,.\, u_i, \sigma^2\right).$$

X is distributed around a mean which is *linearly* according to the values of its parents (Friedman, Linial, Nachman, & Pe'er, 2000). The joint distribution is considered as a multivariate Gaussian , where all the variables in a network have linear Gaussian conditional distributions (Lauritzen & Wermuth, 1989).

c) **Hybrid Network.** Here if the network's structure is represented by a mixture of discrete and continuous variables. According to (Friedman, Linial, Nachman, & Pe'er, 2000) *conditional Gaussian* distributions (Lauritzen & Wermuth, 1989) is used when a continuous variable X has discrete parents. Then a linear Gaussian distribution of X given its continuous parents is represented for each joint assignment to the discrete parents of X (Friedman, Linial, Nachman, & Pe'er, 2000).

Static BNs has an careful restriction that gives the structure of genetic networks, this restriction is because of its acyclicity constraint (Lebre, 2009). This limitation can be solved by providing Dynamic Bayesian network (DBNs) which is used for analyzing gene expression time series by Friedman et al. (Friedman, Murphy, & Russel, 1998)

In DBNs each variable has two time slice $(t \; and \; t + \Delta t)$ [2]. So directed edges means, the edge from nodes at time t to the nodes they effected by the nodes at time $t + \Delta t$ (Needham, Bradford, Bulpitt, & Westhead, 2007). To infer genetic regulatory interactions from microarray data, DBNs have been used (Needham, Bradford, Bulpitt, & Westhead, 2007).

2.2 DYNAMIC BAYESIAN NETWORK

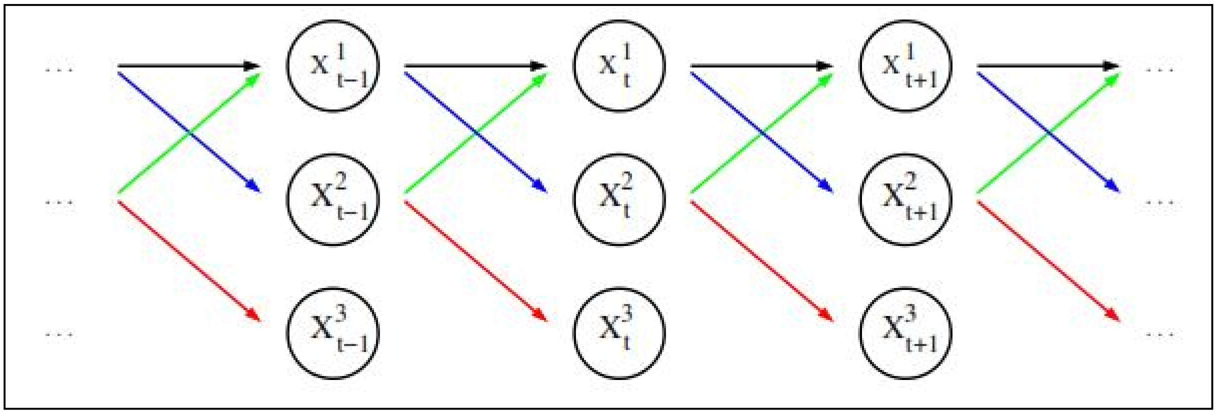
Feedback is an important topic in many biological systems (Needham, Bradford, Bulpitt, & Westhead, 2007). For modeling time series and feedback loops, BNs are absolutely appropriated for these aims (Needham, Bradford, Bulpitt, & Westhead, 2007). If the variables are indexed by time and replicated in the BN, so this mean BNs are used to model time series and feedback loops, these type of networks are called as dynamic Bayesian networks (DBNs) (Needham, Bradford, Bulpitt, & Westhead, 2007). See figure 2 (Lebre, 2009).

Until now many DBN representations that uses different probabilistic models have been used (discrete models (Ong, Glasner, & Page, 2002) (Zou & Conzen, 2005), multivariate autoregressive process (Opgen-Rhein & Strimmer, 2007), State Space or Hidden Markov Models (Perrin, Ralaivola, Mazurie, Bottani, Mallet, & d'Alche Bue, 2003) (Wu, Zhang, & Kusalik, 2004) (Rangel, et al., 2004), nonparametric additive regression model (Imoto, Goto,

& Miyano, 2002) (Imoto, et al., 2003) (Kim, Imoto, & Miyano, 2004) (Sugimoto & Iba, 2004). Kim et al. (Kim, Imoto, & Miyano, 2003) is a review of such models.

So here we will illustrate all the needed conditions for a DBN. For allowing such a DBN representation the existence of a minimal DAG $\mathcal{G}$ will be showed (Lebre, 2009). Then the approximation of $\mathcal{G}$ by q^{th} order conditional dependence DAGs $\mathcal{G}^{(q)}$ is proposed and their probabilistic properties is analyzed by the reduction the dimension of the estimation of the topology of $\mathcal{G}$. Inclusion relationships between the DAGs $\mathcal{G}$ and $\mathcal{G}^{(q)}$ is established from conditions on the topology of $\mathcal{G}$ and the faithfulness assumption (Lebre, 2009) . Then results are used on DAGs $\mathcal{G}^{(q)}$ (Lebre, 2009).

Figure 2.2 X_t^i shows the expression level of gene G^i at time t



Source: Lebre, S. (2009). *Inferring dynamic genetic networks with low order independencies.*

2.3 G1DBN ALGORITHM

The q^{th} order dependence DAGs $\mathcal{G}^{(q)}$ has been recognized in (Lebre, 2009). Here the non-Bayesian inference method for DAG $\mathcal{G}$ providing a DBN representation for process X is used (Lebre, 2009). From Corollary 3 in (Lebre, 2009) q_{max} is assumed to be the maximal number of parents in $\mathcal{G}$. In Corollary 3, inferring $\mathcal{G}$ amounts to inferring $\mathcal{G}^{(qmax)}$ (Lebre, 2009). So there are $\binom{qmax}{p-1}$ potential sets that can guide to conditional independence (Lebre, 2009). In order to develop an inference procedure for $\mathcal{G}$, the true DAG $\mathcal{G}$ is a subgraph of $\mathcal{G}^{(1)}$ (Propositon 6) in (Lebre, 2009). The inference of $\mathcal{G}^{(1)}$ is more faster and more accurate (Lebre, 2009). Then the 2 step-procedure is recognized for DBN inference , which is implemented in a R package 'G1DBN' (Lebre, 2012) freely available from the CRAN.

2.3.1 First Step Of G1DBN (inferring $\mathcal{G}^{(1)}$)

The likelihood of an edge $\left(X_{t-1}^{j}, X_{t}^{i}\right)$ is estimated by calculating the conditional dependence between the variables X_{t-1}^{j} and X_{t}^{i} given any variable X_{t-1}^{k}. The partial regression coefficient $a_{ij|k}$ is considered,

$$X_{t}^{i} = m_{ijk} + a_{ij|k}X_{t-1}^{j} + a_{ik|j}X_{t-1}^{k} + \eta_{t}^{i,j,k},$$

Where the rank of the matrix $\left(X_{t-1}^{j}, X_{t}^{i}\right)_{t\geq2}$ equals 2 and the errors $\{\eta_{t}^{i,j,k}\}_{t\geq2}$ are centered, have same variance and are not correlated.

The conditional dependence between the variables X_{t-1}^{j} and X_{t}^{i} is calculated and given any variable X_{t-1}^{k}, then by testing the null assumption $\mathcal{H}_{0}^{i,j,k}$: "$a_{ij|k} = 0$" . To such purpose, one of the three M-estimators for this coefficient is used: either the familiar Least Square (LS) estimator, the Huber estimator, or Tukey bisquare (or biweight) estimator. The estimates $\acute{\alpha}_{ij|k}$ are computed according to one of these estimators and get the p-value $p_{ij,k}$from the standard significance test as follow :

$$\text{under } (\mathcal{H}_{0}^{i,j,k}: a_{ij|k}=0, \qquad \frac{\acute{\alpha}_{ij|k}}{\sigma(\acute{\alpha}_{ij|k})} \sim t(n-4),$$

Where $t(n-4)$ refers to a student probability distribution with $n-4$ degrees of freedom and $\sigma(\acute{\alpha}_{ij|k})$ is the variance estimates for $\acute{\alpha}_{ij|k}$.

Hence, a score $\mathrm{S_1(i,j)}$ is allocated to each possible edge $\left(\mathrm{X_{t-1}^{j}, X_t^{i}}\right)$ equal to the maximum $\mathrm{Max_{k \neq j}(P_{ijk})}$ of the $\mathrm{p-1}$ computed p-values, which is the best result to first-order conditional independence. It is important to mention that this method does not obtain p-values for the edges but let to order the potential edges of DAG $\mathcal{G}^{(1)}$ according to how similar (likely) they are. The most significant edges for $\mathcal{G}^{(1)}$ means the smallest score . The estimated DAG $\mathcal{G}^{(1)}$ consist of the edges assigned the score below a chosen threshold α_1.

2.3.2 Second Step Of G1DBN

The inferred DAG $\boldsymbol{\mathcal{G}}^{(1)}$ is used as a reduction of the search space. The regression coefficient is denoted by $a_{ij}^{(2)}$ for each pair (i,j) such that the set of edges $(X_{t-1}^{i}, X_t^{i})_{t>1}$ is in $\boldsymbol{\mathcal{G}}^{(1)}$:

$$X_t^i = m_{i\,+} \sum_{j \in pa(X_t^i, \mathcal{G}^{(1)})} a_{ij}^{(2)}\, X_{t-1}^{i} + \eta_t^i,$$

Where the rank of the matrix $(X_{t-1}^{j})_{t \geq 2, j \in pa(X_t^i, \mathcal{G}^{(1)})}$ *is* $|pa(X_t^i, \mathcal{G}^{(1)})|$ *and the errors* $\{\eta_t^i\}_{t \geq 2}$ *are centered, have the same variance, and are not correlated .*

A score $S_2(i,j)$ equal to the p-value $P_{i,j}^{(2)}$ gained from the significance test for each edge of $\mathcal{G}^{(1)}$,

$$under\ (\mathcal{H}_0^{i,j}) : \mathrm{a_{ij}^{(2)}}{=}0, \qquad \frac{\acute{\alpha}_{i,j}^{(2)}}{\sigma(\acute{\alpha}_{i,j}^{(2)})} \sim t\left(n-1-\left|pa\left(X_t^i, \mathcal{G}^{(1)}\right)\right|\right)$$

The score $S_2(i,j) = 1$ are the edges that are not in $\mathcal{G}^{(1)}$. The smallest score means the most significant edges. The estimated DAG consist of the edges assigned the score below a chosen threshold α_2.

The first step of G1DBN results a good estimation of $\mathcal{G}$, this is proved in the Precision-Recall curves in (Lebre, 2009) , also better results can be gained from second step of G1DBN which needs to tune the α_1 and α_2. In the first step of G1DBN α_1 is used for the selection threshold of the edges of $\mathcal{G}^{(1)}$, while α_2 is used for the selection threshold of the edges of $\mathcal{G}$ between the edges of $\mathcal{G}^{(1)}$.

Algorithm1: steps of G1DBN (Lebre, 2009)

```
Choose either LS, Huber or Tukey estimator and set $\alpha_1$ and $\alpha_2$ thresholds.
  inferring $\mathcal{G}^{(1)}$.
For all $i \in P$,
For all $i \in P$,  for all $k \neq j$, compute the p-value $p_{ij|k}$,
$S_1(i,j) = Max_{k \neq j}(P_{ij|k})$.
$E(\mathcal{G}^{(1)}) = \{(X_{t-1}^{j}, X_t^{i})_{t>1}; i,j \in P.\ such\ that\ S_1(i,j) < \alpha_1\}$.
   Step 2: inferring $\mathcal{G}$ from $\mathcal{G}^{(1)}$.
If $N_{pa}^{Max}(\mathcal{G}^{(1)}) \sim n - 1$, choose a higher threshold $\alpha_1$ and go to Step1.

For all $i$ such that $N_{pa}(X_t^i, \mathcal{G}^{(1)}) \geq 1$, copute the p-value $p_{ij}^{(2)}$
$S_2(i,j) = \begin{cases} p_{i,j}^{(2)} for\ all\ i,j \in P\ such\ that\ (X_{t-1}^{j}, X_t^{i})_{t>1} \in \mathcal{G}^{(1)}, \\ 1\ otherwise. \end{cases}$
$E(\mathcal{G}) = \{(X_{t-1}^{j}, X_t^{i})_{t>1}; i \in P, (i,j) \in P\ such\ that\ S_2(i,j) < \alpha_2\}$
```

2.3.3 Choice Of The Threshold

The selection of the threshold is not something easy, specially while utilizing multiple testing. It is difficult to use standard approaches to choose α_1 threshold. Thus a heuristic approach to choose α_1 is used (Butte, Tamayo, Slonim, Golub, & Kohane, 2000). In general, α_1 threshold is chosen after Step 1, where the number of genes have only on parent in DAG $\mathcal{G}^{(1)}$ (Lebre, 2009).

Unlike α_1 threshold, α_2 threshold is provided easier (Lebre, 2009). The usual thresholds are 1%, 5% or 10% or even lower threshold when a low number of edges is needed (Lebre, 2009).

2.4 C3NET ALGORITHM

In this section we illustrate c3net algorithm , its components and an example of its working methods will be introduced.

C3net algorithm is composed of two steps (Altay & Emmert-Streib, 2010). In the first step of c3net the non-significant connections are eliminated to each gene pairs (Lebre, 2009). This can be achieved by testing the statistical significance of pair-wise mutual information (MI) values absorbing resampling methods, which is similar to previous methods, e.g., RN or ARACNE (Butte, Tamayo, Slonim, Golub, & Kohane, 2000) (Margolin, et al., 2006). Mathematically, the mutual information (Cover & Thomas, 1991) of two variables X and Y which are random is described as follow

$$I(X,Y) = \sum_{x \in X} \sum_{x \in Y} p(x,y) log \frac{p(x,yz)}{p(x)p(y)} \qquad (1)$$

The mutual information is calculated from the data by using a suitable estimator allowing a close approximation of the theoretical value of the population (Altay & Emmert-Streib, 2010). Started form a fully connected matrix C, with C_{ij} = 1 for all i, $j \in V$ and a zero matrix A, all pair-wise mutual information values I_{ij}, $i, j \in V$ are comprehensively tested, and $C_{ij} = C_{ji}$ = 0 is set if the null hypothesis H_0: I_{ij} = 0 cannot be rejected, for a given significance level α (Altay & Emmert-Streib, 2010). In the second step of C3NET, first the neighborhood N_s is determined, for all genes $i \in V$ (Altay & Emmert-Streib, 2010). To define the neighborhood of gene , N_s $N_s = \{j : C_{ij} = 1 \; and \; j \neq i\}$ is used (Altay & Emmert-Streib, 2010). For this purpose the auxiliary connectivity matrix C is introduced (Altay & Emmert-Streib, 2010). The connection of each gene to its neighborhood that has the maximum mutual information value is determined from N_s and I (Altay & Emmert-Streib, 2010). This connection is determined by

$$J_c(i) = argmax\{I_{ij}\} \qquad (2)$$

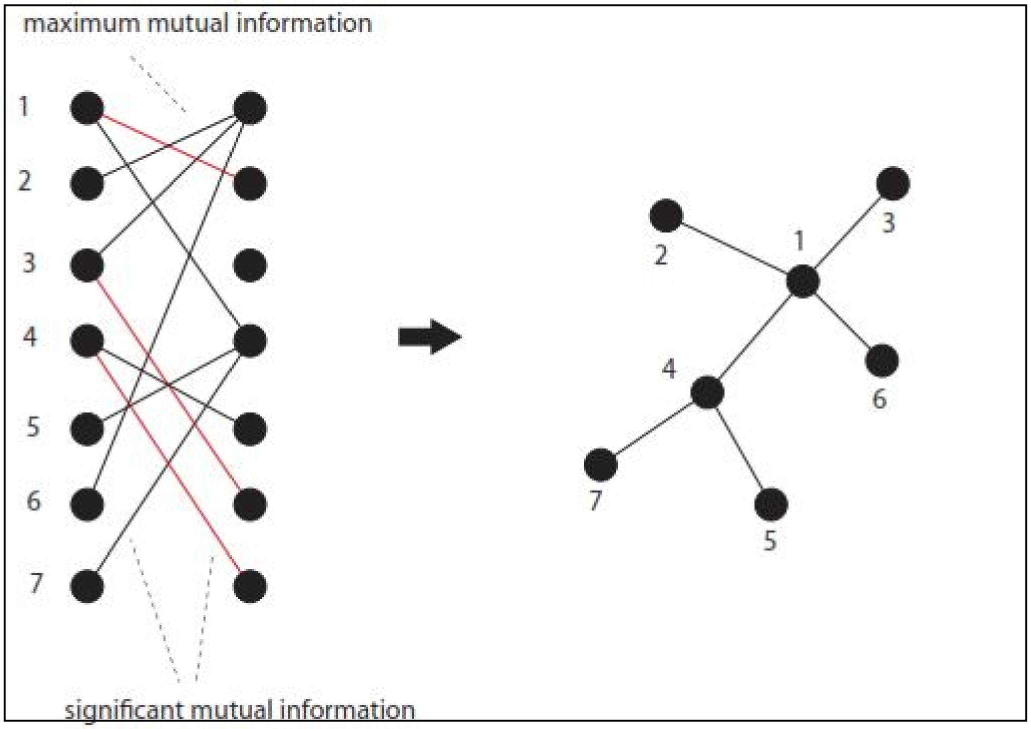
If all mutual information values I_{ij} for $j \in V$ were non-significant so $N_s(i) \neq \emptyset$ so an index is not assigned to $j_c(i)$ but the empty set is assigned (Altay & Emmert-Streib, 2010). From this information the adjacency matrix A of the estimated undirected network by setting $A_{ij_c(i)} = A_{jc(i)i} = 1$ if $j_c(i)$ is set to an index (Altay & Emmert-Streib, 2010). All other entries is set to zero or remain zero (Altay & Emmert-Streib, 2010). The principle steps of the method are explained in algorithm 1 (Altay & Emmert-Streib, 2010). Finally, a gene can have relation with more than one gene. This is indicated with a simple example composed of four genes. Fig. 1 explain the example .

Algorithm 2 Steps of inference algorithm C3NET as shown in (Altay & Emmert-Streib, 2010).

```
$A$: initiate adjacency matrix, $A_{ij} = 0$ for all $i, j \in V$
$C$: initiate connectivity matrix, $C_{ij} = 1$ for all $i, j \in V$
estimate mutual information $I_{ij}$ for all $i, j \in V$
**repeat**
    Set $C_{ij} = 0$ if $I_{ij} = 0$ is not statistically significant (hypothesis test)
**until** all pairs $i \neq j$ are tested
**for all** $i \in V$ **do**
    **if** $N_s(i) = \{j : C_{ij} = 1 \; and \; j \neq i\}$
```

```
  if $N_s(i) = \emptyset$
    $j_c(i) = argmax_{j\in N_s(i)\{I_{ij}\}}$
  else
    $j_i(i) = \emptyset$
  endif
end for
for all $i \in V$ do
  if $j_c(\boldsymbol{i}) = \emptyset$
    $A_{ij_c(i)} = A_{j_c(i)i} = 1$
  endif
end for
return adjacency matrix $A$
```

Figure 2.3 Fundamental mechanism of C3NET.The red and black edges are the significant edges. The edges in black are the maximum mutual information at the left hand side.



Source: Altay, G., & Emmert-Streib, F. (2011). Structural influence of gene newtworks on their inference: analysis of C3NET. *Biology Direct*

For example there are mutual information values *I* and its corresponding connectivity matrix *C*, as a result of hypotheses tests, as follow taken from (Altay & Emmert-Streib, 2010):

$$I = \begin{pmatrix} 1.0 & 0.7 & 0.9 & 0.8 \\ 0.7 & 1.0 & 0.6 & 0.5 \\ 0.9 & 0.6 & 1.0 & 0.1 \\ 0.8 & 0.5 & 0.1 & 1.0 \end{pmatrix}, C = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 0 \\ 1 & 1 & 1 & 0 \\ 1 & 0 & 0 & 1 \end{pmatrix} \quad (3)$$

Connection with neighboring genes are specified for each of the four genes with maximum mutual information which is also statistically significant, in $j_c = (3, 1, 1, 1)$ is resulted (Altay & Emmert-Streib, 2010). Mutual information values that are not statistically significant are set to zero in the matrix *C* (Altay & Emmert-Streib, 2010). From j_c an auxiliary matrix can be determined,

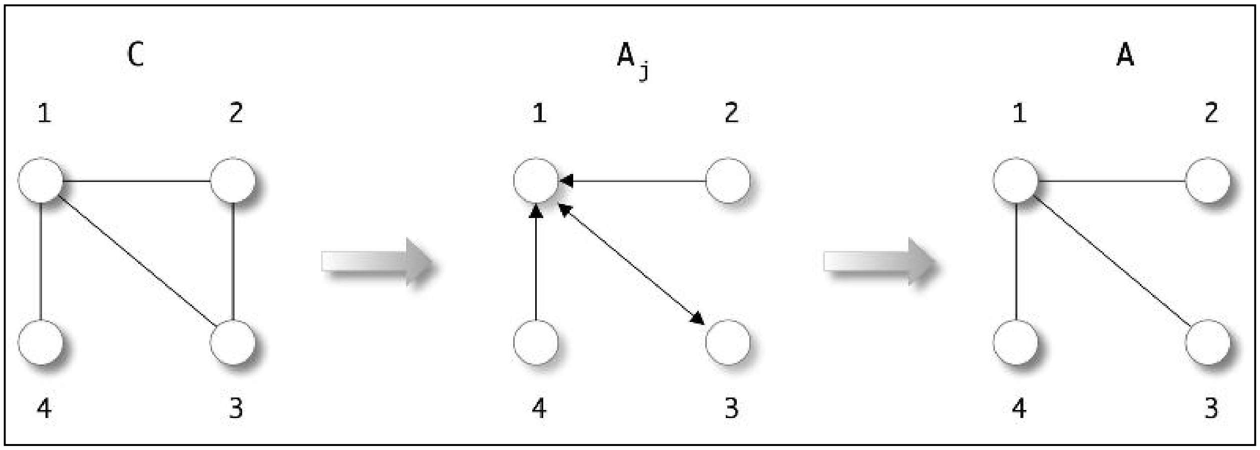
$$A_j = \begin{pmatrix} 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{pmatrix} \qquad (4)$$

Which contains the exact edges added by each node (Altay & Emmert-Streib, 2010). MI information dose not support directional information, because its argument's symmetry, so the resulting adjacency matrix *A* is a symmetric adjacency matrix (Altay & Emmert-Streib, 2010).

$$A = \begin{pmatrix} 0 & 1 & 1 & 1 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{pmatrix} \qquad (5)$$

From Fig. 2 which is taken from (Altay & Emmert-Streib, 2010) we can see inferred network provide by adjacency matrix *A* is star-like and gene 1 is connected to 3 other genes (Altay & Emmert-Streib, 2010).

Figure 2.4 C3NET algorithm



Source: Altay, G., & Emmert-Streib, F. (2010). Inferring the conservative causal core of gene regulatory networks. *BMC System Biology.*

The computational complexity of C3NET is $O(n^2)$ since matrices which has since of $n \times n$ can enter C3NET procedure, this according to the pseudo code of C3NET algorithm in Algorithm 1 (Altay & Emmert-Streib, 2010). We know that the knowledge about biological regulatory networks are still not completed, so simulated data is used because their true regulatory network is known (Altay & Emmert-Streib, 2010). This provides a good and detailed analysis.

The simulation study is complemented with biological expression data to explain that the assumptions made for our simulations are realistic enough to estimate these results to biological data sets (Altay & Emmert-Streib, 2010). F-score is used to provide the performance of an inference algorithm, $= 2pr/(p + r)$ (Altay & Emmert-Streib, 2010) . Here the precision, $p = TP/(TP + FP)$, and recall, $r = \frac{Tp}{TP+FN}$, Is the function of true positive number (TP), false positive (FP) and false negative (FN) edges in an assumed network (Altay & Emmert-Streib, 2010). The capabilities of an inference algorithm the way in (Meyer, Kontos, & Bonternpi, 2007) is followed, which obtain an optimal cutoff value I_0 for the mutual information values by maximizing the F-score (Altay & Emmert-Streib, 2010) (Meyer, Kontos, & Bonternpi, 2007). Two biological networks are used in C3NET simulation study (Altay & Emmert-Streib, 2010), which they are subnetworks of the transcriptional regulatory network (TRN) of E. coli (Shen-Orr, Orr, Milo, Mangan, & Alon, 2002) (Ma, Kumar, Ditges, Gunzer, Buer, & Zeng, 2004) and Yeast (Guelzim, Bottani, Bourgine, & Kepes, 2002). These subnetworks were randomly sampled with the *neighbor addition* method from these TRNs using SynTReN (Van den Bulche, et al., 2006). SynTReN is a generator of synthetic gene expression data which is used for design and analysis of structure learning algorithms (Van den Bulche, et al., 2006). The networks were consisted of *n* = 100 nodes (genes) (Altay & Emmert-Streib, 2010).

Synthetic expression data (including noise) mimicking the mRNA concentration in steady-state condition by using non-linear transfer functions based on Michaelis-Menten and Hill enzyme kinetic equations (Fersht, 1985) (Mendes, Sha, & Ye, 2003) were generated with SynTReN (Van den Bulche, et al., 2006). For C3NET (Altay & Emmert-Streib, 2010) simulations ensemble approach is used (Emmert-Streib & Altay, 2010) (Altay & Emmert-Streib, 2010). Due to estimate the mutual information values for the synthetic data sets first, copula-transform is applied to the data (Altay & Emmert-Streib, 2010). After that a parametric Gaussian estimator is applied to estimate MI values (Altay & Emmert-Streib, 2010), as illustrated in (Meyer, Kontos, & Bonternpi, 2007) and (Olsen, Meyer, & Bontempi, 2009), the MI values are estimated by

$$I(X,Y) = \left(\frac{1}{2}\right)\log\left(\frac{\sigma^2X\sigma^2Y}{|C|}\right) \qquad (6)$$

Here σ^2X and σ^2Y is the variance of X respectively Y and |C| is the determinant of the covariance matrix (Altay & Emmert-Streib, 2010). (Milller-Madow, Shrikage or Schurmann-Grassberger (Meyer, Kontos, & Bonternpi, 2007) (Meyer, Lafitte, & Bontempi, 2008) they are estimator which can be used in C3NET algorithm but did not provide a better performance, so the fastest estimator for (Altay & Emmert-Streib, 2010) simulations is used. *E.* coli data set is the biological expression is used in (Altay & Emmert-Streib, 2010) which taken from (Faith, et al., 2007). Due to obtain a reference network that can be used to provide the performance an inference algorithm a curated network is assumed mostly depends on the RegulonDB database (Gama-Castro, et al., 2008).

Implementation of C3NET: Using the R package (Altay & Emmert-Streib, 2010) (Altay & Emmert-Streib, 2011)

C3NET is made usable for bioglogists by implementing a R package called *c3net* (Altay & Emmert-Streib, 2011). The software package c3net is available from the web site https://r-forge.r-project.org/projects/c3net and from the CRAN package repository.

The principle working mechanism of the c3net package is demonstrated by providing an example data set (Altay & Emmert-Streib, 2010). In c3net package the *data(expdata)* and *data(truenet)* commands are used to call the data set and the true network which are loaded in R (Altay & Emmert-Streib, 2010). Here the *expdata* and *truenet* are the variables of data set and true network respectively (Altay & Emmert-Streib, 2011).

There is a function of the package *c3net* which takes the data set as input and outputs the inferred network (Altay & Emmert-Streib, 2011). The detail of the function is: *c3net(dataset, alpha = 0.01, methodstep1= "cutoff", cutoff MI = 0, MTCmethod = "BH', itnum = 5, network = FALSE)* (Altay & Emmert-Streib, 2011). Here *dataset* and *alpha* are the data set and user defined significance level α respectively (Altay & Emmert-Streib, 2010). The method *methodstep1,* user can set three different options, {*"cutoff", "MTC", "justp"*}, this is for eliminating nonsignificant edges (Altay & Emmert-Streib, 2011). It uses parameter *cutoffMI* if *methodstep1 = "cutoff",* needs a numerical value which is used as cutoff value to eliminate nonsignificant MI value of edges in Step 1 of C3NET (Altay & Emmert-Streib, 2010). A multiple testing correction (MTC) method is used in Step 1 of C3NET if *methodstep1 = "MTC"* (Altay & Emmert-Streib, 2010). In this situation, a MTC method require to be specified by the dependent parameter *MTCmethod* (e.g. *MTCmethod* = "BH") (Altay & Emmert-Streib, 2010). Different methods of MTC are available which are *"BH", "bonferroni", "BY", "hochberg", "holm", "hommel"* (Altay & Emmert-Streib, 2010). To provide a null distubution and *alpha* the statistical significance level, the *itnum* required to be assign to specify the number of iterations (Altay & Emmert-Streib, 2010). Only *alpha* and *itnum* need to be assigned if *methodstep1 = "justp",*and the elimination in Step1 is done according to the p-values and the significance level α only (Altay & Emmert-Streib, 2010).

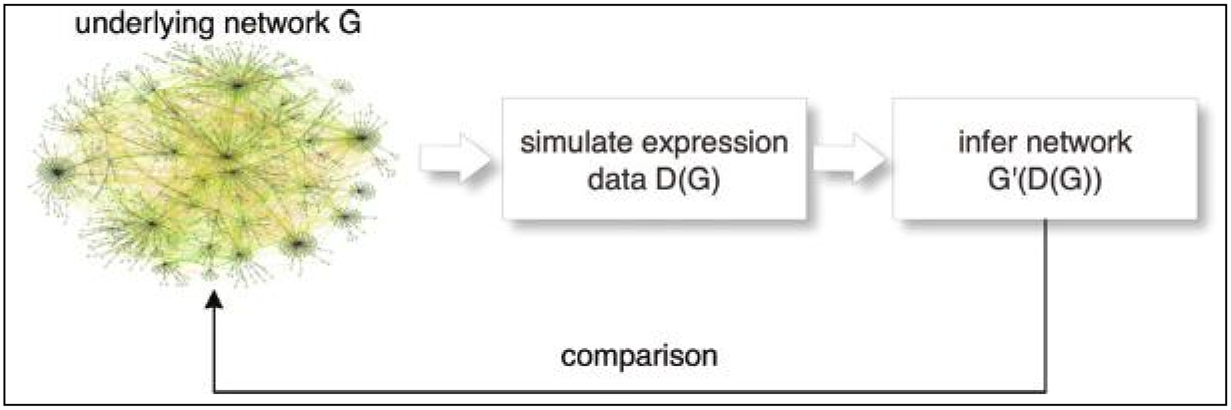
The c3net package has the plotting option of the inferred network by using the *igraph* package (Altay & Emmert-Streib, 2010). This can be published by assigning the parameter *network* in *c3net* function to *TRUE* (e.g *c3net(expdata, network = TRUE*) (Altay & Emmert-Streib, 2011). In *c3net* package there is another important function to establish the performance of the inference called *checknet* (Altay & Emmert-Streib, 2011).This done by executing *checknet(net, truenet)* (Altay & Emmert-Streib, 2011). The output of the *checknet* function is as follows: (prescision = 0.96, F-score = 0.34, recall = 0.21, TP = 181, FP = 6, FN = 683) (Altay & Emmert-Streib, 2010).

2.5 DATA SET PREPARATION

The synthetic network we use in our approach represents subnetworks of the transcriptional regulatory network (TRN) of *E.* coli (Shen-Orr, Orr, Milo, Mangan, & Alon, 2002) (Ma, Kumar, Ditges, Gunzer, Buer, & Zeng, 2004) and DAG (Pearl, 1988), we call them reference networks. These (reference) subnetworks were randomly sampled with the *neighbor addition* method from these TRNs using SynTReN (Van den Bulche, et al., 2006). With SynTReN (Van den Bulche, et al., 2006) we sampled two simulated data sets from reference networks. From these data sets we inferred a network according to our approach. In order to calculate the performance of our inference approach we use (reference) networks.

As we noticed that the information about biological regulatory networks are not being completed (Altay & Emmert-Streib, 2010), so we use simulated (reference) data since these data's underlying (true) regulatory network is known exactly. This let us make a detailed and accurate analysis.

Figure 2.5 llustrates data set preparation.



Source: Altay, G., & Emmert-Streib, F. (2011). Structural influence of gene newtworks on their inference: analysis of C3NET. *Biology Direct*

We plot the reference networks see figure 6 and figure 7**.** The structure of reference network that used in c3net is $n \times n$ matrix, since the inferred network of c3net is undirected (Altay & Emmert-Streib, 2010). Actually the structure of reference network is composed of three column, first column is for the prediction of gene the second is for the target gene and the third column is for the score of direction (Lebre, 2012).

Figure 2.6 Reference network for DAG 100x100 sample.

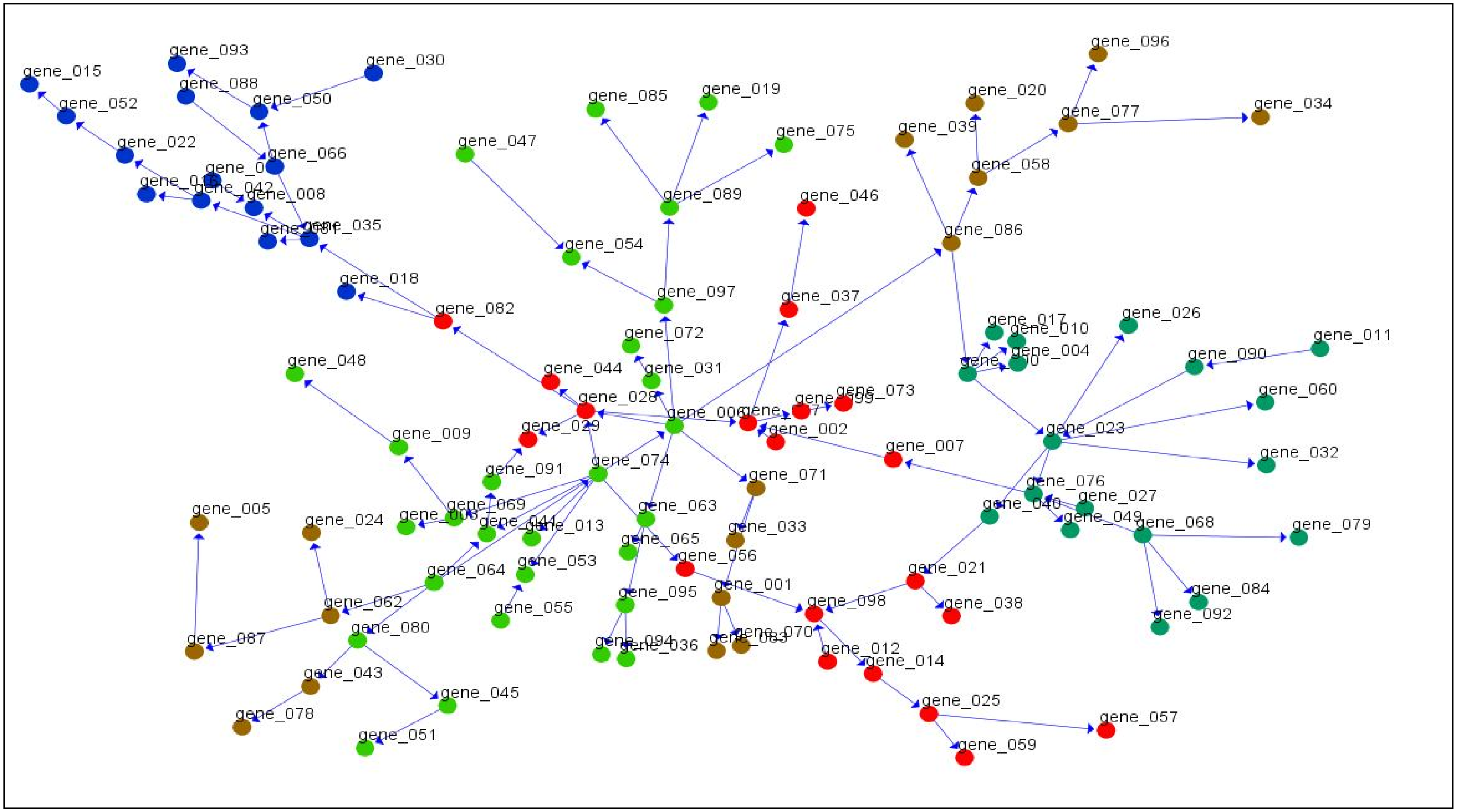
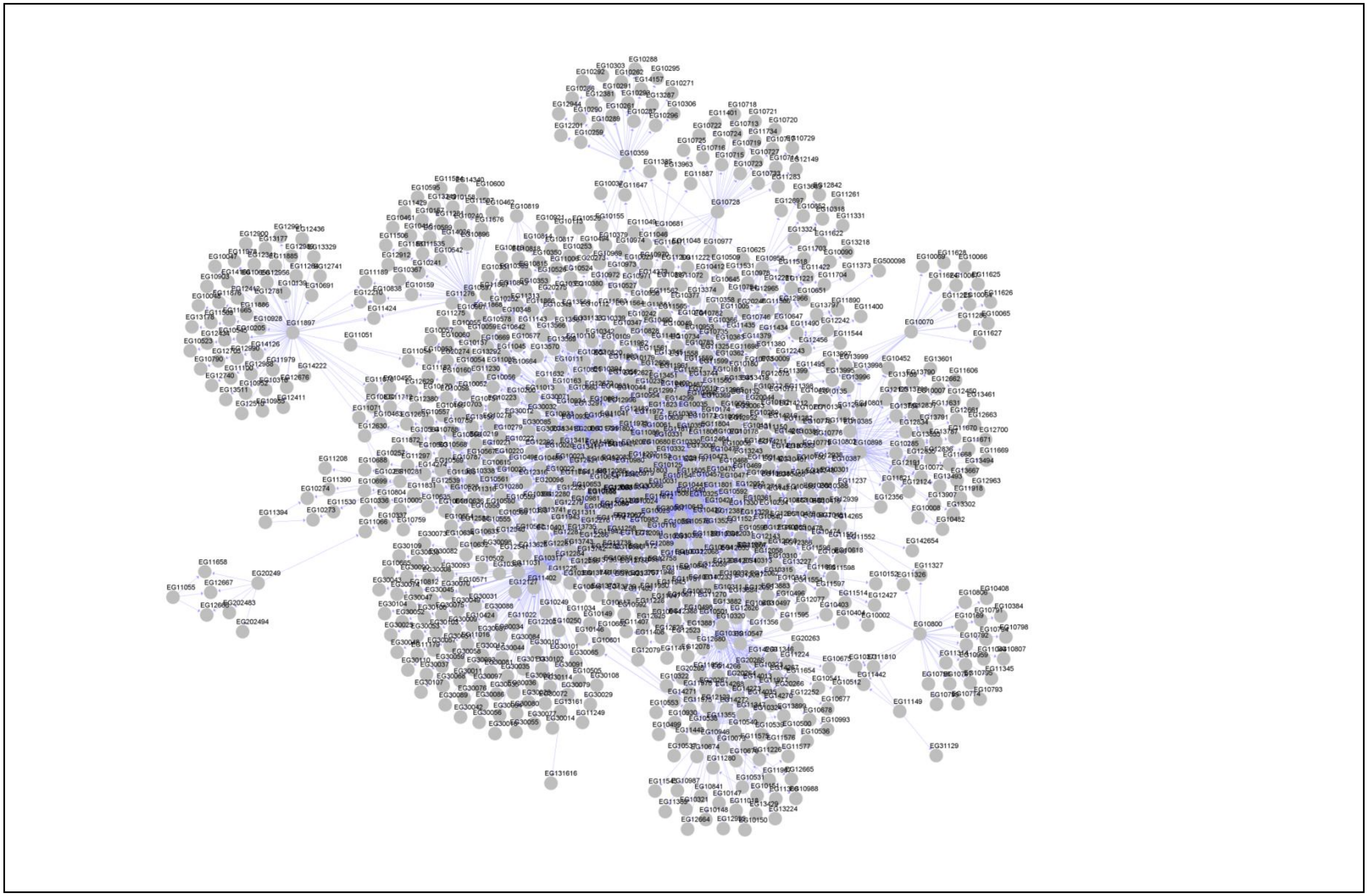


Figure 2.7 Reference directed network for *E*.coli



3. FINDINGS

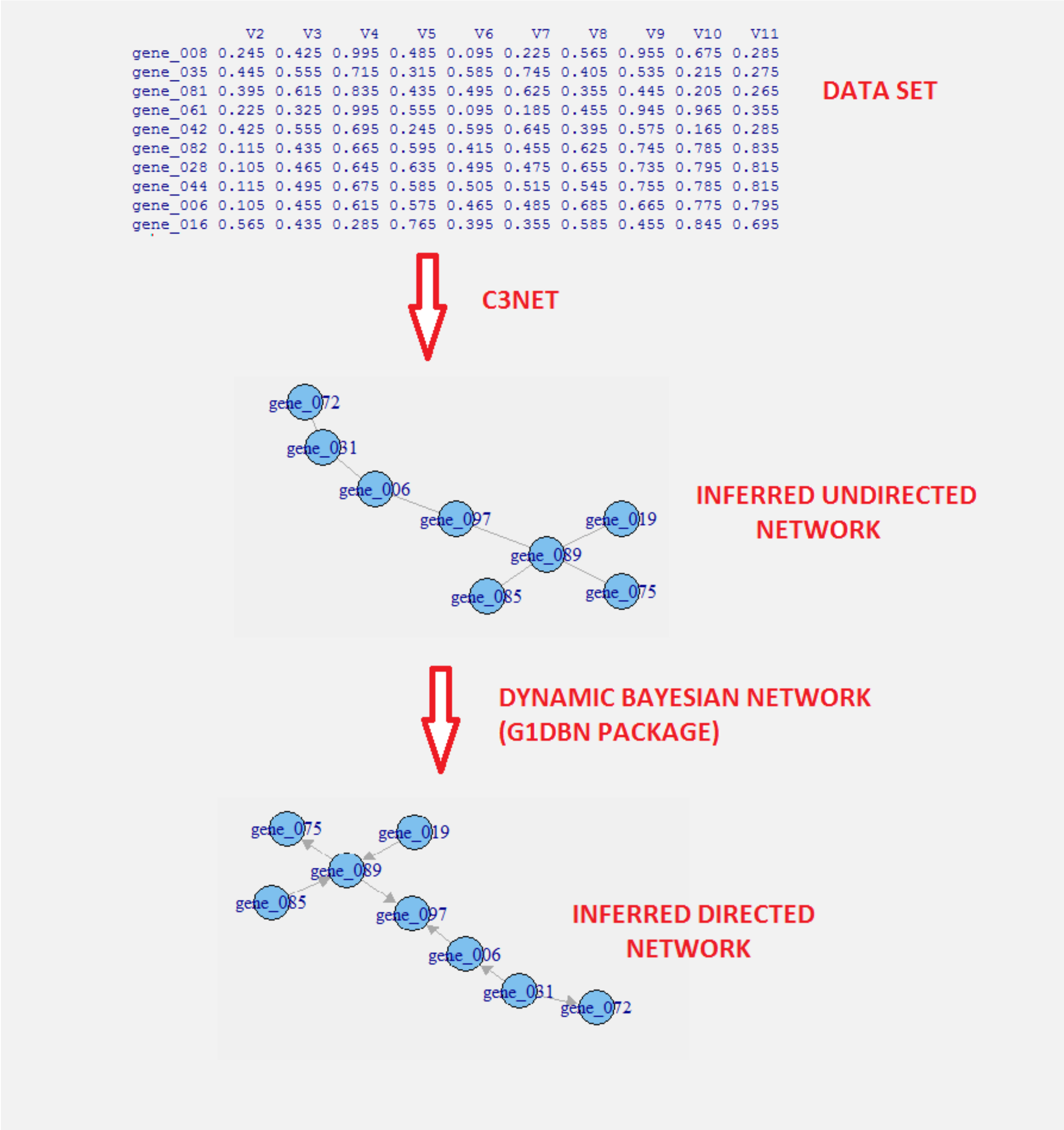
3.1 OUR APPROACH

In this section we introduce our approach, describe its components and show its mechanism. Our approach is composed of two steps, in the first step we apply c3net algorithm (Altay & Emmert-Streib, 2011) to the synthetic data set, returning a symmetric mutual information matrix $n \times n$ (Altay & Emmert-Streib, 2011). The non-zero elements in the returned matrix show undirected edges between variables which are statistically significant (tested in the first step of C3NET) (Altay & Emmert-Streib, 2011). We illustrated the C3NET algorithm in data and method section. In the second step we take the inferred network of C3NET algorithm which is the $n \times n$ matrix and apply the G1DBN (Lebre, 2009) to each pair of that has interaction to find the direction between them. The inference procedures implemented in R package 'G1DBN' and C3NET is available from the CRAN archive. G1DBN is a package performs Dynamic Bayesian Network inference (Lebre, 2009) as we mentioned. The aim of applying the dynamic Bayesian network to the output of C3NET algorithm is to make the undirected edges between variables which are statistically significant to directed edges. Then evaluate the performance of the inferred networks. Figure 8 illustrates the mechanism of our approach.

Algorithm3: Steps of our approach

```
1: $A$: adjacency matrix $A$ from C3NET algorithm 1.
2: for all pair $A_{ij} = 1$ do
3: G1DBN algorithm 2.
4: returns estimated DAG $\mathcal{G}^{(1)}$: matrix $S_1(i,j)$
```

Figure 3.1 Two steps of our approach.



```
              V2    V3    V4    V5    V6    V7    V8    V9   V10   V11
gene_008 0.245 0.425 0.995 0.485 0.095 0.225 0.565 0.955 0.675 0.285
gene_035 0.445 0.555 0.715 0.315 0.585 0.745 0.405 0.535 0.215 0.275
gene_081 0.395 0.615 0.835 0.435 0.495 0.625 0.355 0.445 0.205 0.265
gene_061 0.225 0.325 0.995 0.555 0.095 0.185 0.455 0.945 0.965 0.355
gene_042 0.425 0.555 0.695 0.245 0.595 0.645 0.395 0.575 0.165 0.285
gene_082 0.115 0.435 0.665 0.595 0.415 0.455 0.625 0.745 0.785 0.835
gene_028 0.105 0.465 0.645 0.635 0.495 0.475 0.655 0.735 0.795 0.815
gene_044 0.115 0.495 0.675 0.585 0.505 0.515 0.545 0.755 0.785 0.815
gene_006 0.105 0.455 0.615 0.575 0.465 0.485 0.685 0.665 0.775 0.795
gene_016 0.565 0.435 0.285 0.765 0.395 0.355 0.585 0.455 0.845 0.695
```

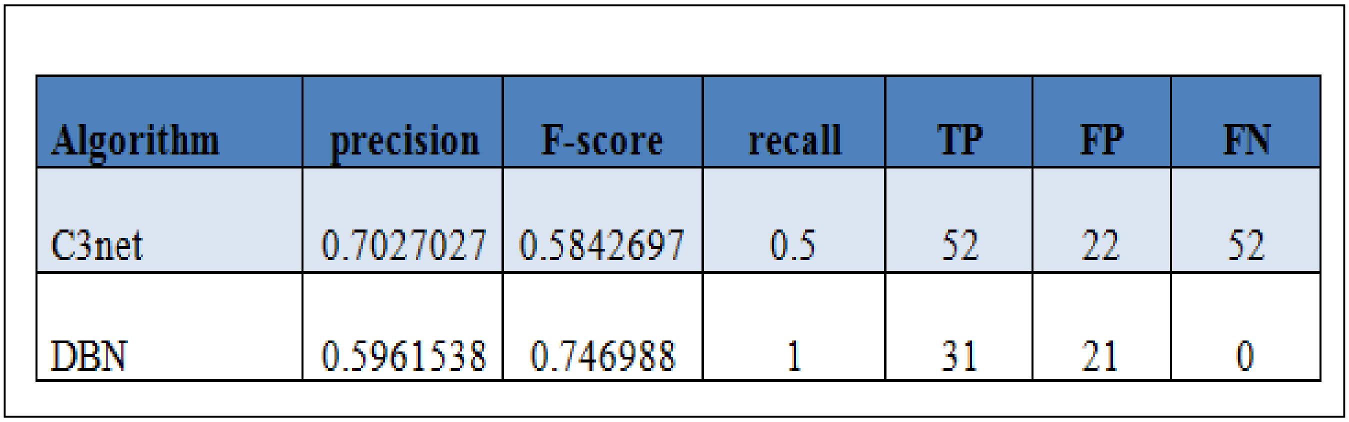
3.2 APPLYING OUR APPROACH

In this section we apply our approach to the simulated data sets and illustrate the results with examples.
For the both synthetic data set *E.* coli (Shen-Orr, Orr, Milo, Mangan, & Alon, 2002) and DAG (Pearl,1988), we apply first step of our approach, which is the C3NET algorithm by c3net package which is available in CRAN. Then we generate the reference undirected network to evaluate the performance of the output of the C3NET (Altay & Emmert-Streib, 2010) algorithm see the table 1 and table 2. Then we apply second step of our approach which is Dynamic Bayesian Network by G1DBN (Lebre, 2009), since the algorithm of the DBN is complex, we apply G1DBN to a pair of gene instead to all of the dataset at the same time. If we apply to all of the data set at the same time, it will take more time. So we apply the G1DBN only to the pair of gene which has interaction. The knowledge of interaction is from the output of the c3net algorithm (Altay & Emmert-Streib, 2010).

For example, when we apply c3net algorithm to the simulated data set of DAG (Shen-Orr, Orr, Milo, Mangan, & Alon, 2002), there were 74 undirected interactions. As you see in the table 1 there are 52 interactions which are TP and 22 interactions which are FP which is equal to 74 interaction. Also we have 52 FN interactions which are the interactions available in the reference network but not available in the inferred network after c3net. So we apply the G1DBN to each pair of the interactions only see algorithm 3. Here we have two performance evaluation one for undirected network, second for directed network. The F-score is 0.5842697, recall is 0.5 and precision is 0.7027027 for undirected network. After the second step of our approach we evaluate the performance of the inferred directed network as in the table 1. We have 31 TP , 21 FP, and 0 FN. The F-score is equal to 0.7, recall is 1 and precision is 0.5961.

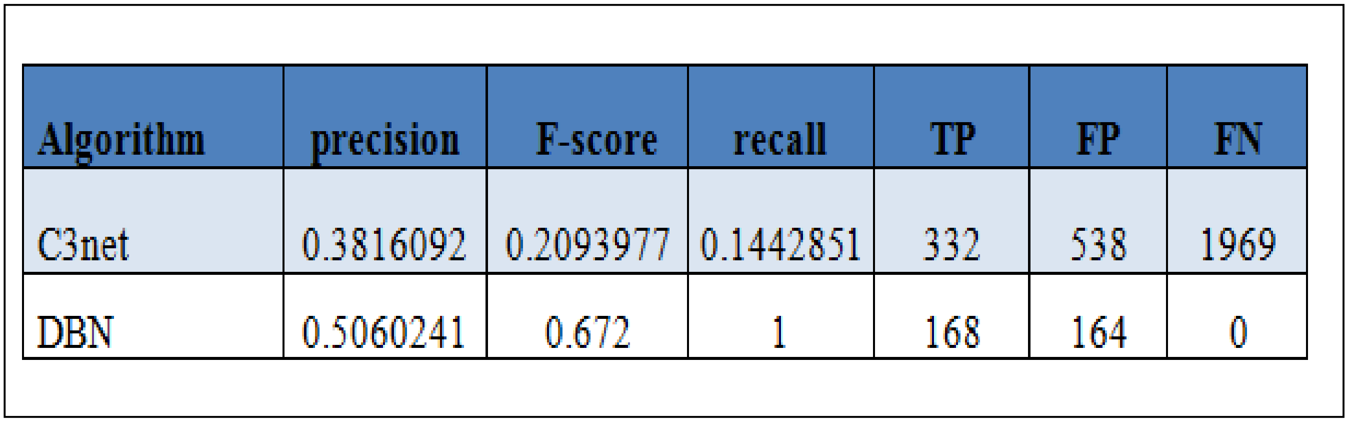
We do the same steps to the second data set see table 2. Then we plot both undirected and directed inferred networks see figure 9 and figure 10. In figure 9 it is the inferred undirected network with 74 interactions where the interactions are undirected, in figure 10 the inferred network is directed after applying our approach, the red blue edges are the 31 TP 21 FP directed edges respectively. We apply the same steps to the simulated dataset of *E.coli* (Shen-Orr, Orr, Milo, Mangan, & Alon, 2002) see figure 11,12. Unlike the data set of DAG (Pearl,1988), the *E.coli* data set is large which have 1000 genes. Actually our approach is very flexible for such large dataset.

Table 3.1 Performance evaluation of the DAG (Pearl, 1988) data set for c3net and G1DBN algorithm



Algorithm	precision	F-score	recall	TP	FP	FN
C3net	0.7027027	0.5842697	0.5	52	22	52
DBN	0.5961538	0.746988	1	31	21	0

Table 3.1 Performance evaluation of the data set *E.* coli (Shen-Orr, Orr, Milo, Mangan, & Alon, 2002) for c3net and G1DBN algorithm



Algorithm	precision	F-score	recall	TP	FP	FN
C3net	0.3816092	0.2093977	0.1442851	332	538	1969
DBN	0.5060241	0.672	1	168	164	0

Figure 3.2 Undirected inferred network from DAG, after applying C3NET algorithm

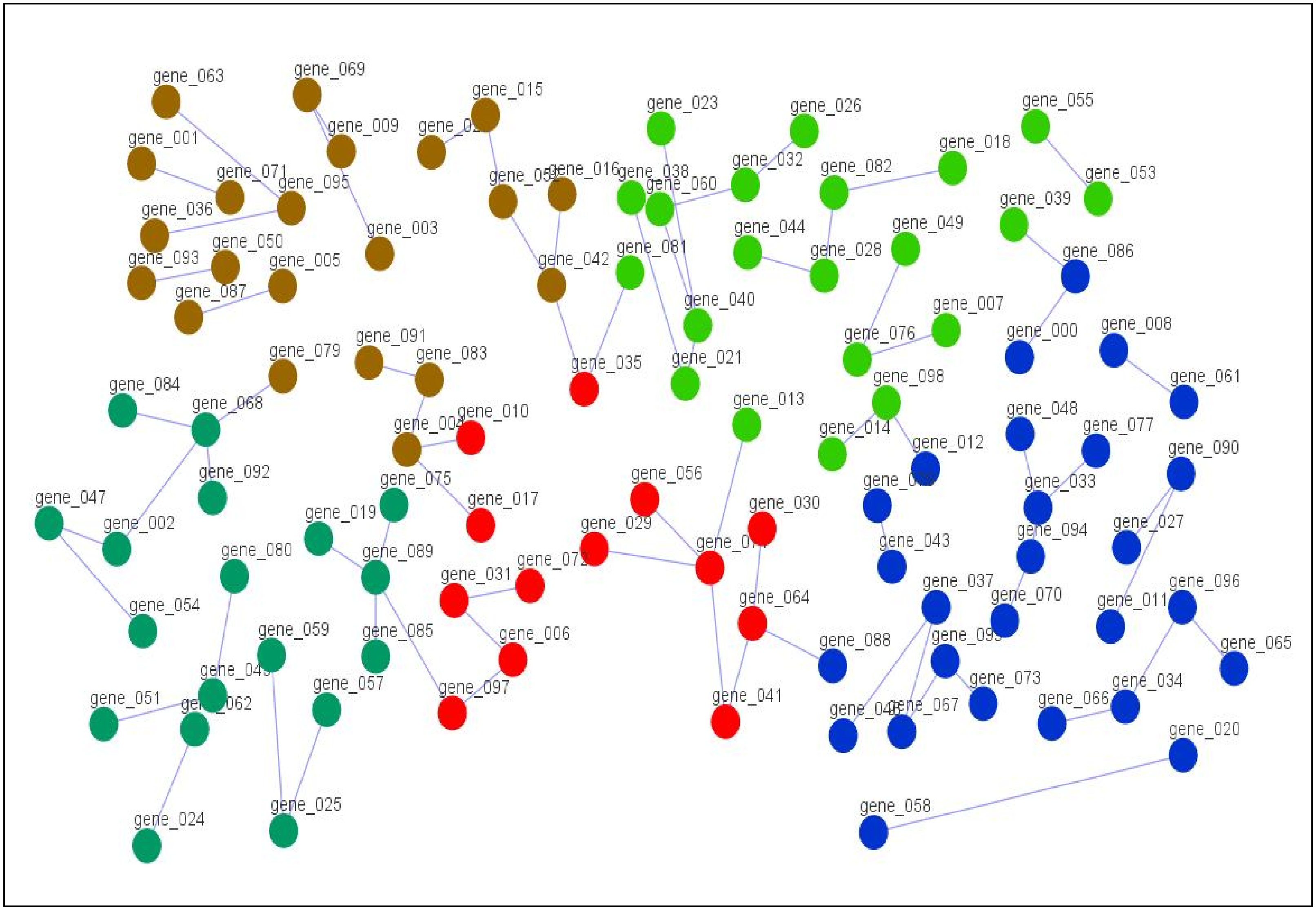


Figure 3.3 Directed inferred network from DAG after applying G1DBN algorithm by our approach, the nodes with the red label and edge are the true positive, and the blue ones are the false positive. Here TP=31, FP=21

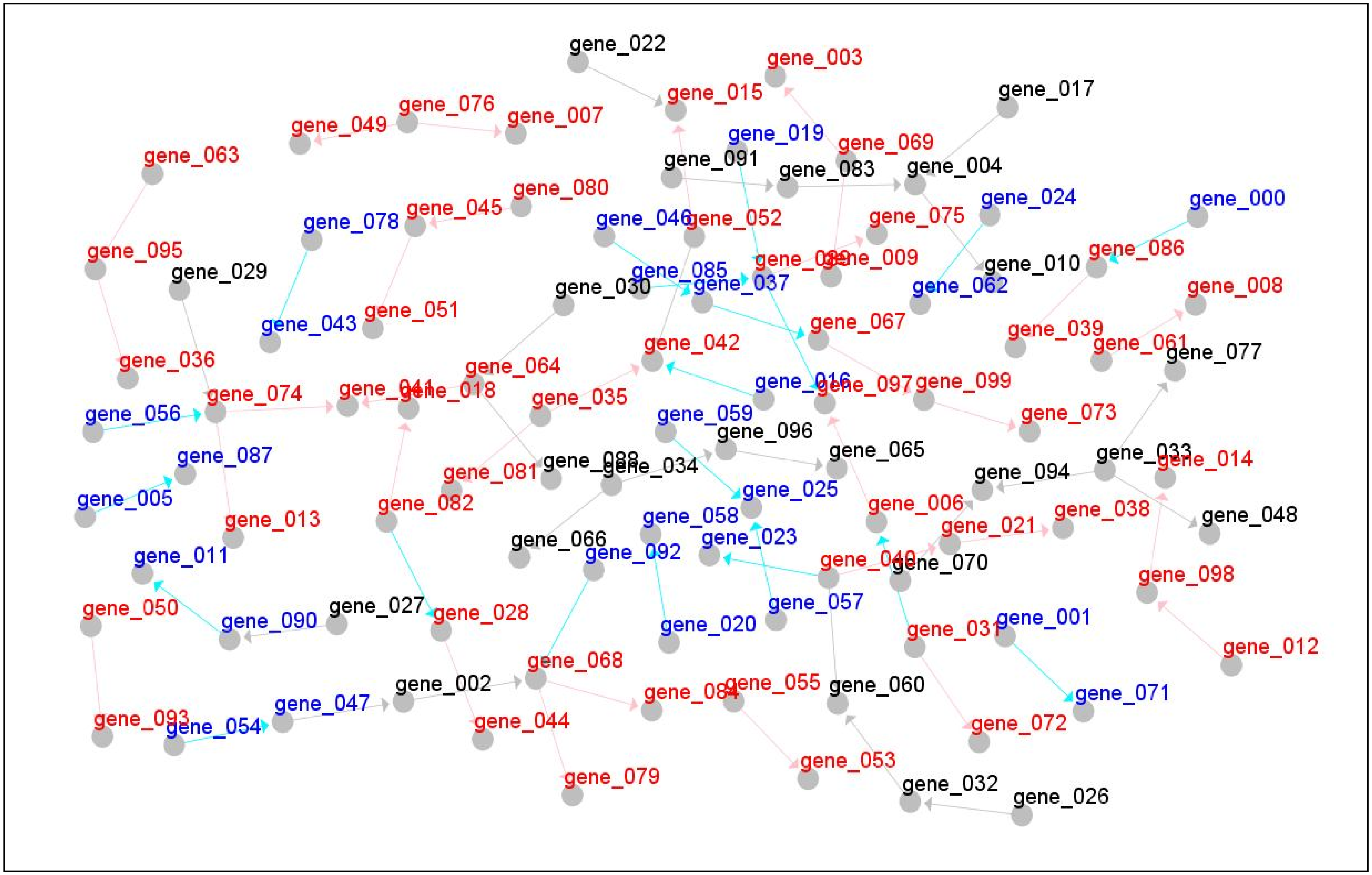


Figure 3.4 Undirected inferred network from *E*.coli , after applying C3NET algorithm

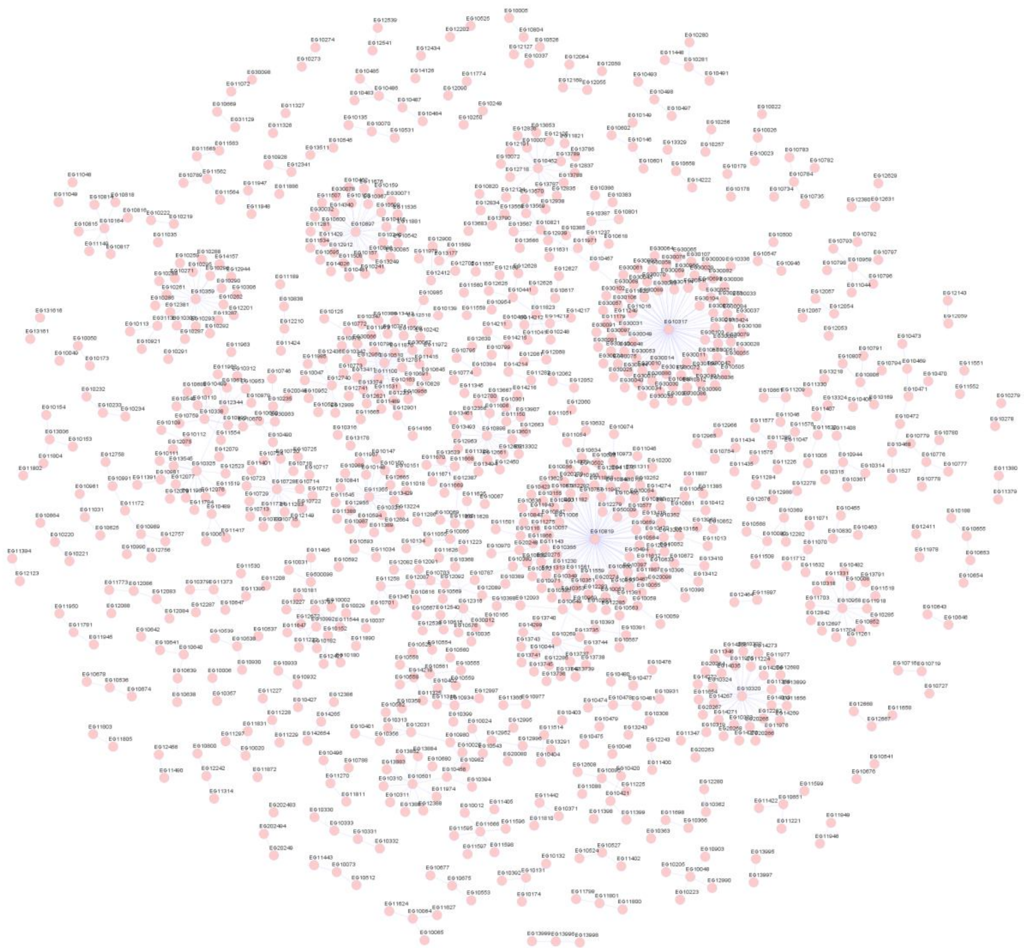
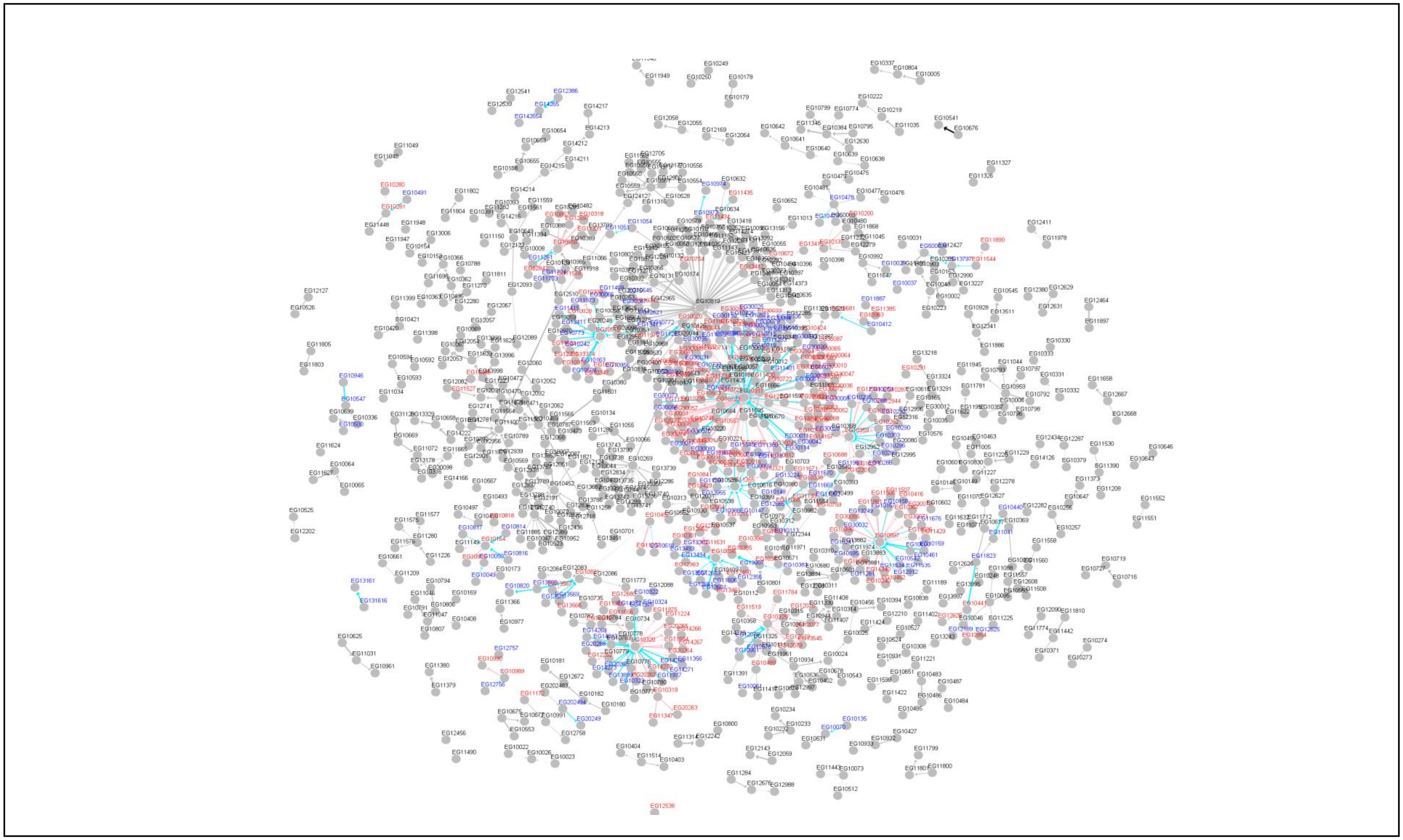


Figure 3.5 Directed inferred network from *E.*coli after applying G1DBN algorithm by our approach, the nodes with the red label and edge are the true positive, and the blue ones are the false positive. Here TP=168, FP=16



4.CONCLUSION

GRNI methods are widely studied in order to infer causal regulatory networks. ARACNE, CLR, MRNET, RN and C3NET are well-known inference methods that are frequently used.

In our study we merge between two algorithms which they are C3NET and G1DBN. Both of them are used for inferring causal interactions between genes. The aim of our study is to convert the undirected inferred network of C3NET to directed inferred network. For this purpose we applied Dynamic Bayesian Network by G1DBN algorithm.

Applying Dynamic Bayesian Network to a large data set is complex and take time, but our approach solves this problem in two steps. In the first step we decrease the probability of the gene interactions by C3NET algorithm, then in the second step we apply the Dynamic Bayesian Network to each pair of nodes not to whole inferred network. For example we applied our approach to the expression data of *E.coli* .

In first step of our approach we applied C3NET algorithm to the 1000x2000 expression data set of *E.coli* we obtained 870 interactions, 332 of the interactions were TP. Then in the second step we applied Dynamic Bayesian Network to each pair of inferred network (870 interactions) by G1DBN. We find 168 TP directed edges without taking time.

Although our approach has been used for inferring causal interactions between genes, it may be used in another field and applications such as causal relations among covariates, since the requirements for the data are moderate.

REFERENCES

Books:

Cover, T., & Thomas, J. 1991. *Information Theory*. New York: John Wiley & Sons. *Inc* .

Cox, D. R., & Wermuth, N. 1996. *Multivariate dependencies: Models, analysis and interpretation*. Chapman and Hall.

Friedman, N., Murphy, K., & Russel, S. 1998. *Learning the structure of Dynamic probabilistic networks.* USA: Morgan Kaufmann.

Fersht, A. 1985. *Enzyme structure and mechanism.* New York: W. H. Freeman and Company.

Pearl, J. 1988. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference.* SF, CA, USA: Morgan Kaufmann Publishers.

Holmes, D., & Jain, L. 2008. *Innovations in Bayesian Networks.* New York: Springer-Verlag.

Periodicals:

Altay, G., & Emmert-Streib, F. 2011. Infering large-scale gene networks with C3NET.

Altay, G., & Emmert-Streib, F. 2010. Inferring the conservative causal core of gene regulatory networks. *BMC System Biology* .

Altay, G., & Emmert-Streib, F. 2010. Revealing differences in gene network inference algorithms on the network-level by ensemble methods. *Bioinformatics* , 26(14):1738-44.

Altay, G., & Emmert-Streib, F. 2011. Structural influence of gene newtworks on their inference: analysis of C3NET. *Biology Direct* .

Ding, C., & Peng, H. 2005. Minimum redundancy feature selection from mivroarray gene expression data. *Journal of Bioinformatics and Computational Biology* , 3(2):185-205.

Dudoit, S., Shaffer, J., & Boldrick, J. 2003. Multiple hypothesis testing in microarray experiments. *Statistical Science* , 18:71-103.

Emmert-Streib, F., & Altay, G. 2010. Local network-based measures to assess the inferability of different regulatory network. *IET Syst Biol* , 4(4):277-88.

Emmert-Streib, F., & Dehmer, M. 2010. Medical Biostatistics for Complex Diseases. *Weinheim: Wiley-Blackwell* .

Emmert-Streib, F., 2011. Networks for Systems Biology: Conceptual Connection of Data and Function. *IET Systems Biology* , 5(3):185-207.

Faith, J., Hayete, B., Thaden, J., Mogno, I., Wierzbowski, J., Cottarel, G., et al. (2007). Large-Scale Mapping and Validation of Escherichia Profiles. *PloS Biol* , 5.

Friedman, N., Linial, M., Nachman, I., & Pe'er, D. 2000. Using Bayesian networks to analyse expression data. *Journal of Computational Biology* , 7(3-4):601-620.

Imoto, S., Kim, S., Goto, T., Aburatani, S., Tashiro, K., Kuhara, S., et al. 2003. Bayesian network and nonparametric heteroscedastic regression. *Journal of Bioinformatics* , 2:231–252.

Kim, S., Imoto, S., & Miyano, S. 2003. Inferring gene networks from time series microarray data using dynamic Bayesian networks. *Bioinformatics* , 4(3):228.

Margolin, A., Nemenman, I., Basso, K., Wiggins, C., Stolovitzky, G., Dalla Favera, R., et al. 2006. ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC Bioinformatics* , 7:57.

Mendes, P., Sha, W., & Ye, K. 2003. Artificial gene networks for objective comparison of analysis algorithms. *Bioinformatics* , 19:122-129.

Meyer, P., Kontos, K., & Bonternpi, G. 2007. Information-theoretic inference of large transcriptional regulatory networks. *EUROSIP journal on bioinformatics ans systems biology* .

Meyer, P., Lafitte, F., & Bontempi, G. 2008. minet: A R/Bioconductor Pachage for Inferring Large Transcriptional Networks Using Mutual Information. *BMC Bioinformatics* , 9:461.

Olsen, C., Meyer, P., & Bontempi, G. 2009. On the Impact of Entropy Estimator in Transcriptional Regulatory Network Inference. *EURASIP Journal on Bioinformatics and Seystems Biology* .

Ong, I. M., Glasner, J. D., & Page, D. 2002. Modelling regulatory pathways in e. coli from time series expression profiles. *Bioinformatics*.

Opgen-Rhein, R., & Strimmer, K. 2007. Learning causal networks from systems biology time course data: and effective model selection procedure for the vector autoregressive process. *BMC Bioinformatics* , 8(Suppl. 2):S3.

Perrin, B. E., Ralaivola, L., Mazurie, A., Bottani, S., Mallet, J., & d'Alche Bue, F. (2003). Gene networks inference using dynamic Bayesian networks. *Bioinformatics* .

Rangel, C., Angus, J., Ghahramani, Z., Lioumi, M., Sotheran, E., Gaiba, A., et al. (2004). Modelling t-cell activation using gene expression profiling and state-space models. *Bioinformatics* .

Schafer, J., & Strimmer, K. 2005. An empirical bayes approach to inferring large-scale gene association networks. *Bioinformatics* , 21:754-764.

Scutari, M. (2010). Learning Bayesian Networks with the bnlearn R Package. *Journal of Statistical Software*

Steuer, R., Kurths, J., Daub, C., Weise, J., & Selbig, J. 2002. The mutual information: Detecting and evaluating dependencies between variables. *Bioinformatics* , 18(2):5231-240.

Steuer, R., Kurths, J., Fiehn, O., & Wechwerth, W. 2003. Observing and interpreting correlations in metabolomic networks. *Bioinformatics* , 19(8):1019-1026.

Toh, H., & Horimoto, K. 2002. Inference of a genetic network by a combined approach of cluster analysis and graphical gaussian modeling. *Bioinformatics* , 18:287-297.

Van den Bulche, T., Van Leemput, K., Naudts, B., van Remortel, P., Ma, H., Verschoren, A., et al. 2006. SynTReN: a generator of synthetic gene expression data for design and analysis of structure learning algorithms. *BMC Bioinformatics* , 7:43.

Wang, J., Myklebost, O., & Hovig, E. 2003. Mgraph: graphical models for microarray data analysis. *Bioinformatics* .

Xing, B., & van der Laan, M. 2005. A causal inference approach for constructing transcriptional regulatory networks. *Bioinformatics 2005* , 21(21):4007-4013.

Zou, M., & Conzen, S. D. 2005. A new dynamic Bayesian network (DBN) approach for identifying gene regulatory networks from time course microarray data. *Bioinformatics* , 21(1):71-79.

Other Publications:

Acid, S., de Campos, L., Fenandes-Luna, J., Rodriguez, J., & Salcedo, J. 2004. A Comparison of Learning Algorithms for Bayesian Networks: A Case Study Base on Data from An Emergency Medical Service. *Artificial Intelligence i Medicine* .

Beaumont, M., & Rannala, B. 2004. The Bayesian revolution in genetics. *Nat Rev Genet* , 5: 251-261.

Bradford, J., Needham, C., Bulpitt, A., & Westhead, D. 2006. Insights int protein-protein interfaces using a Bayesian network prediction method. *JMol Biol* , 362: 365-386.

Butte, A., & Kohane, I. 2000. Mutual information relevance networks: Functional genomic clustering using pairwise entropy measurements. *Pacific Symposioum on Biocomputing* , 5:415-426.

Butte, A., Tamayo, P., Slonim, D., Golub, T. R., & Kohane, I. S. 2000. Discovering functional relationships between RNA expression and chemotheraputic susceptibility using relevance networks. *PNAS* .

Friedman, N. 2004. Inferring cellular networks using probabilistic graphical models. *Science* , 303: 799-805.

Gama-Castro, S., Jimenez-Jacinto, V., Peralta-Gil, M., Santos-Zavaleta, A., Penaloza-Spinola, M., Contreras-Moreira, B., et al. 2008. RgulonDB (version 6.0): gene regulation model of Escherichia coli K-12 beyond transcription, active (experimental) annotated promoters and Textpresso navigation. *Nucl Acids Res* , 36(suppl 1):D120-124.

Guelzim, N., Bottani, S., Bourgine, P., & Kepes, F. 2002. Topological and causal structure of the yeast transcriptional regulatory network. *Nature Genetics* , 31:60-63.

Imoto, S., Goto, T., & Miyano, S. 2002. Estimation of genetic networks and functional structures between genes by using Bayesian networks and non-parametric regression. *Pacific Symposium on Biocomputing 7* , 175–186.

Kim, S., Imoto, S., & Miyano, S. 2004. Dynamic Bayesian network and nonparametric regression for nonlinear modeling of gene networks from time. *Biosystems* , 75(1-3):57–65.

Kraskov, A., Stagbaur, H., & Grassberger, P. 2004. Estimating mutual information. *Phys Rev E* , 69(6):066138.

Lauritzen, S. L. 1996. Graphical models. *Oxford Statistical Science Series* .

Lauritzen, S., & Wermuth, N. 1989. Graphical models for associations between variables, some of which are qualitative and some quantitative. *Annals of Statistics* , 17: 31-57.

Lebre, S. 2012. A package performing Dynamic Bayesian Network inference.

Lebre, S. 2009. Inferring dynamic genetic networks with low order independencies.

Li, H., & Gui, J. 2006. Gradient directed regularization for sparse Gaussian concentration graphs, with applications to inference of genetic networks. *Biostatistics* , 7(2):302-317.

Li, W. 1990. Mutual information functions versus correlation functions. *Journal of Statistical Physics* , 60(5-6):823-837.

Ma, H., Kumar, B., Ditges, U., Gunzer, F., Buer, J., & Zeng, A. 2004. An extended transcriptional regulatory network of Escherichia coli and analysis of its hierarchical structure and network motifs. *Nucleic Acids Res* , 32:6643-6649.

Needham, C. J., Bradford, J. R., Bulpitt, A. J., & Westhead, B. D. 2007. A Primer on Learning in Bayesian Networks for Computational Biology. *PLoS Coputational Biology* .

Sachs, K., Perez, O., Pe'er, D., Lauffenburger, D., & Nolan, G. 2005. Causal protein-signaling networks derived from multiparameter single-cell data. *Science* , 308: 523-529.

Schafer, J., & Strimmer, K. 2005. A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. *Statistical Applications in Genetics and Molecular Biology* , 4(32).

Scutari, M. 2013. Bayesian network structure learning, parameter learning and inference.

Shen-Orr, S., Orr, S., Milo, R., Mangan, S., & Alon, U. 2002. Network motifs in the transcriptional regulatory network of Escherichia coli. *Nat Genes* , 31:64-68.

Speed, T. 2003. Statistical Anaylsis of Gene Expression Microarray Data.

Sugimoto, N., & Iba, H. 2004. Inference of gene regulatory networks by means of dynamic differential Bayesian networks and nonparametric regression. *Genome Informatics* , 15(2):121–130.

Toh, H., & Horimoto, K. 2002. System for automatically inferring a genetic network from expression profiles. *J. Biol. Physics* .

Tourassi, G., Frederick, E., Markey, M., & Floyd, C. 2001. Application of the mutual information criterion for feature selection in computer-aided diagnosis. *Medical Physics* , 28(12):2394-2402.

Vidal, M. 2009. A unifying view of 21st century systems biology. *FEBS Letters* , 583(24):3891-3894.

Waddell, P. J., & Kishino, H. 2000. Cluster inference methods and graphical models evaluated on nci60 microarray gene expression data. *Genome Informatics* , 11:129-140.

Waddell, P. J., & Kishino, H. 2000. Correspondence analysis of genes and tissue types and finding genetics links from microarray data. *Genome Informatics* , 11:83-95.

Watkinson, J., Liang, K., Wang, X., Zheng, T., & Anastassiou, D. 2009. Inference of regulatory gene interactions from expression data using three-way mutual information. *Ann N Y Acad Sci* , 1158:302-13.

Wu, F. X., Zhang, W. J., & Kusalik, A. J. 2004. Modelling gene expression from microarray expression data with state-space equations. *In Pacific Symposium on Biocomputing* , 581-592.

Wu, X., Ye, Y., & Subramanian, K. R. 2003. Interactive analysis of gene interactions using graphical gaussian model. *ACM SIGKDD Workshop on Data Mining in Bioinformatics* , 3:63-69.