

CS 791 Graduate Seminar

Fall 2018

Inference of Gene Regulatory Network Based on Local Bayesian Networks

Kiran Teja Sarvamthota

Hari Chandana Chintha

ABSTRACT:

The inference of gene regulatory networks (GRNs) from expression data can mine the direct regulations among genes and gain deep insights into biological processes at a network level. Local Bayesian network (LBN), to infer GRNs from gene expression data by using the network decomposition strategy and false-positive edge elimination scheme. LBN algorithm uses conditional mutual information (CMI) to construct an initial network or GRN, which is decomposed into a number of local networks or GRNs. BN method is employed to generate a series of local BNs by selecting the k-nearest neighbors of each gene as its candidate regulatory genes, which significantly reduces the exponential search space from all possible GRN structures. Bayesian networks not only effectively reduce the computational cost of BN due to much smaller sizes of local GRNs, but also identify the directions of the regulations.

INTRODUCTION:

Accurately inferring GRN is of great importance and also an essential task to understand the biological activity from signal emulsion to metabolic dynamics, prioritize potential drug targets of various diseases, devise effective therapeutics, and discover the novel pathways. Identifying the GRNs with experimental methods is usually time-consuming, tedious and expensive, and sometimes lack of reproducibility. Overall, these model-based methods can provide us a deeper understanding of the system's behaviors at a network level and can also infer the directions of regulations in the network. However, these methods are parameters-dependent and time-consuming, which makes them difficult to deal with large-scale networks. To overcome these limitations of BN, MI and CMI, in this paper, we discussed a novel local Bayesian network (LBN) algorithm.

PROBLEM STATEMENT:

The model-based methods can provide us a deeper understanding of the system's behaviors at a network level and can also infer the directions of regulations in the network. However, these methods are parameters-dependent and time-consuming, which makes them difficult to deal with large-scale networks. To overcome these limitations of BN, MI and CMI, in this paper, we discuss a novel local Bayesian network (LBN) algorithm to reconstruct GRNs from gene expression data by making use of their advantages, i.e., infer the directed network with less false-positive edges and with high computational efficiency.

PROBLEM SIGNIFICANCE:

Bayesian network (BN) methods cannot handle large-scale networks due to their high computational complexity, while information theory-based methods cannot identify the directions of regulatory interactions and also suffer from false positive/negative problems. To overcome the limitations, in this work we discuss a novel algorithm, namely local Bayesian network (LBN), to infer GRNs from gene expression data by using the network decomposition strategy and false-positive edge elimination scheme.

PROPOSED METHOD:

To overcome these limitations of BN, MI and CMI, in this paper, we study a novel local Bayesian network (LBN) algorithm to reconstruct GRNs from gene expression data by making use of their advantages, i.e., infer the directed network with less false-positive edges and with high computational efficiency.

LBN algorithm mainly consists of five distinct elements: i) CMI is first employed to construct an initial network, i.e., local networks or GRNs, according to the nearest relationship among genes in the network with k-nearest neighbor (kNN) method. ii) For these local networks or GRNs, BN method is used to identify their regulatory relationships with directions, generating a series of local BNs which are integrated into a candidate GRN GB. iii) CMI is applied to remove the false positive edges in GB, forming a tentative GRN GC. iv) According to the relationships of kNN among genes in the network, the tentative GRN (GC) is further decomposed into a series of smaller sub-networks or local networks, in which BN method is implemented to delete the false regulatory relationships. v) The final network or GRN GF is inferred by iteratively performing BN and CMI with kNN decomposition until the topological structure of the tentative network GC does not change.

EXISTING METHODS:

Recently, information theory-based methods are widely used for reconstructing GRNs, such as mutual information (MI) and conditional mutual information (CMI). These approaches are assumption-free methods, measuring unknown, non-linear and complex associations rather than linear-correlations between genes, and addressing the problem of intense computation for parameters. Thus, they can be used to infer large-scale GRNs. However, MI-based methods overestimate the regulation relationships to some extent and fail to distinguish indirect regulators from direct ones, thereby leading to possible false positives. Although CMI-based

methods are able to separate the direct regulations from the indirect ones, they cannot derive the directions of regulations in the network and also tend to underestimate the regulation strength in some cases.

IMPLEMENTATION:

MI and CMI: mutual information (MI) and conditional mutual information (CMI) are widely implied to inferring GRN. MI can be used to measure the degree of independence between two genes X_i and X_j , but it tends to overestimate the regulation strengths between genes (i.e., false positive problem). On the other hand, CMI measures the conditional dependency between two genes X_i and X_j given other gene X_k , which can quantify the undirected regulation.

For discrete variables X and Y , MI is defined as

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

where $p(x, y)$ is the joint probability distribution of X and Y , and $p(x)$ and $p(y)$ are the marginal probability distributions of X and Y , respectively; $H(X)$ and $H(Y)$ are the entropies of X and Y , respectively; and $H(X, Y)$ is the joint entropy of X and Y .

CMI between two variables X and Y given variable Z is defined as

$$\begin{aligned} CMI(X, Y|Z) &= \sum_{x \in X, y \in Y, z \in Z} p(x, y, z) \log \frac{p(x, y, z)}{p(x|z)p(y|z)} \\ &= H(X, Z) + H(Y, Z) - H(Z) - H(X, Y, Z) \end{aligned}$$

where $p(x, y, z)$, $p(x|z)$ and $p(y|z)$ are conditional probability distributions, and $H(X, Z)$, $H(Y, Z)$, and $H(X, Y, Z)$ are the joint entropies.

The hypothesis of Gaussian distribution for gene expression data, the entropy can be estimated by the following Gaussian kernel probability density function

$$P(X_i) = \frac{1}{N} \sum_{j=1}^N \frac{1}{(2\pi)^{n/2} |C|^{n/2}} \exp\left(-\frac{1}{2} (X_j - X_i)^T C^{-1} (X_j - X_i)\right)$$

where C is the covariance matrix of variable X , $|C|$ is the determinant of the matrix, N is the number of samples and n is the number of variables (genes) in C .

The entropy of variable X can be denoted as:

$$H(X) = \log[(2\pi e)^{n/2}|C|^{1/2}] = \frac{1}{2}\log[(2\pi e)^n|C|]$$

BAYESIAN NETWORKS: A Bayesian network (BN) is a graphical model of the probabilistic relationships among a set of random variable $X = \{X_1, X_2, \dots, X_i, \dots, X_n\}$,

$$P(X_1, X_2, \dots, X_n) = \prod_{X_i \in X} P(X_i | Pa(X_i))$$

which is a directed acyclic graph G .

where $Pa(X_i)$ is the set of parents of node X_i in graph G .

We first construct the undirected network with CMI method, and decompose the network into a series of sub-networks in which the central node just is linked with its k nearest neighbors (or nodes).

k-nearest neighbor: In a graph $G(V, E)$, V represents a set of nodes and E represents edges between nodes. The k closest neighbors of each node are selected according their shortest path distance in the graph structure. That is, the k -nearest neighbor (kNN) of node V_i consists of a set of nodes whose shortest path to the node V_i is k

We used the k -nearest neighbors of each node to decompose a large-scale network to form a series of local Bayesian networks. For each local Bayesian network, the Bayesian network inference method is used to remove the false positive edges. For a large-scale network, we show that it can actually achieve a high accuracy even with the first- and second-nearest neighbors of each node

LBN ALGORITHM: Given an expression dataset with n genes and N samples, a novel algorithm (called LBN) was developed to infer its underlying GRN.

LBN is composed of four main parts:

Step 1: Construct the initial network by CMI. I

Step 2: Decompose GMI into n sub-networks or local networks by kNN.

Step 3: Construct local BNs by estimating the gene regulations and integrate local BNs into a candidate network.

Step 4: Construct tentative network by eliminating the redundant regulations by CMI.

Step 5: Decompose GC into N smaller networks or local networks.

Then we have the final network or GRN by iteratively performing CMI and BN with kNN methods. Numerical computations show that our LBN method can infer the final GRN after iterating 10–20 times.

EXPERIMENTS AND RESULTS:

DATASET:

Toy Dataset: Data provided with the reference of the paper.

Simulated Dataset: Data derived from DREAM challenge are used to evaluate the algorithm. DREAM challenge gives a series of gene expression datasets with noise and gold benchmark networks, which were selected from source networks of real species.

EVALUATION:

In order to validate our algorithm, the true positive rate (TPR), false positive rate (FPR), false discovery rate (FDR), positive predictive value (PPV), overall accuracy (ACC), F-score measure and Matthews correlation coefficient (MCC) are used to evaluate the performance of our LBN and other algorithms. These metrics are defined as follows:

$$TPR = TP / (TP + FN), FPR = FP / (FP + TN),$$

$$FDR = FP / (TP + FP), PPV = TP / (TP + FP),$$

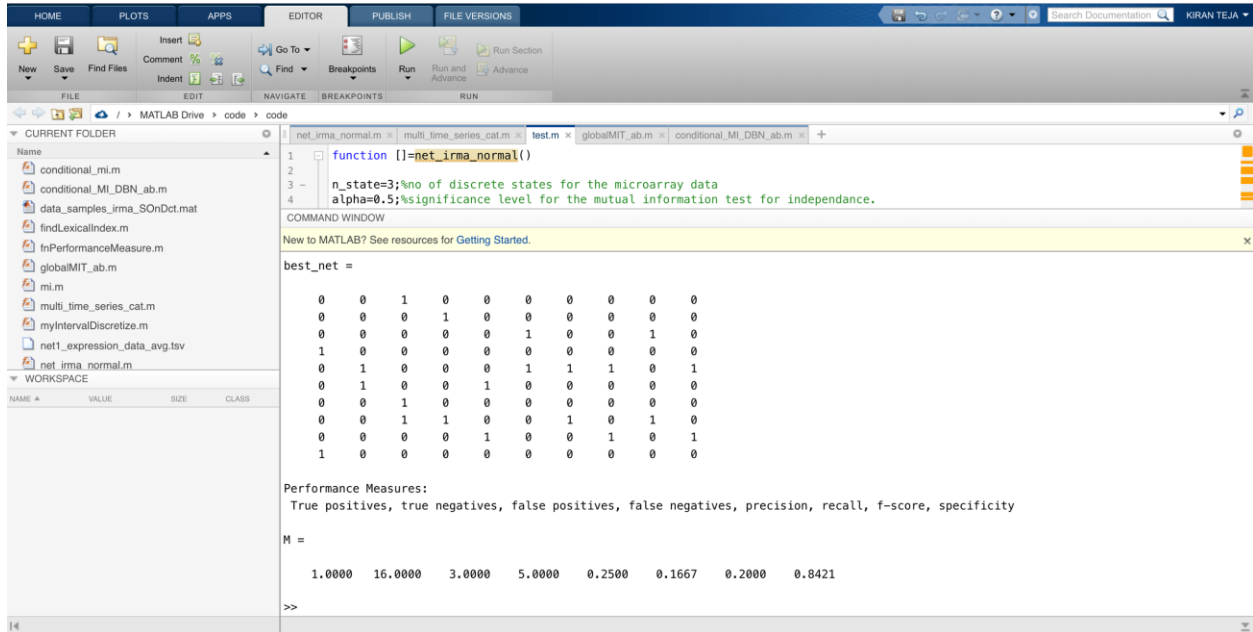
$$ACC = (TP + TN) / (TP + FP + TN + FN),$$

$$F = 2PPV \times TPR / (PPV + TPR),$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

RESULTS:

Best network for 10 Genes:



The screenshot shows the MATLAB Editor with the following code and output:

```
function []=net_irma_normal()
n_state=3;%no of discrete states for the microarray data
alpha=0.5;%significance level for the mutual information test for independence.
```

COMMAND WINDOW

New to MATLAB? See resources for Getting Started.

best_net =

0	0	1	0	0	0	0	0	0	0
0	0	0	1	0	0	0	0	0	0
0	0	0	0	0	1	0	0	1	0
1	0	0	0	0	0	0	0	0	0
0	1	0	0	0	1	1	1	0	1
0	1	0	0	1	0	0	0	0	0
0	0	1	0	0	0	0	0	0	0
0	0	1	1	0	0	1	0	1	0
0	0	0	0	1	0	0	1	0	1
1	0	0	0	0	0	0	0	0	0

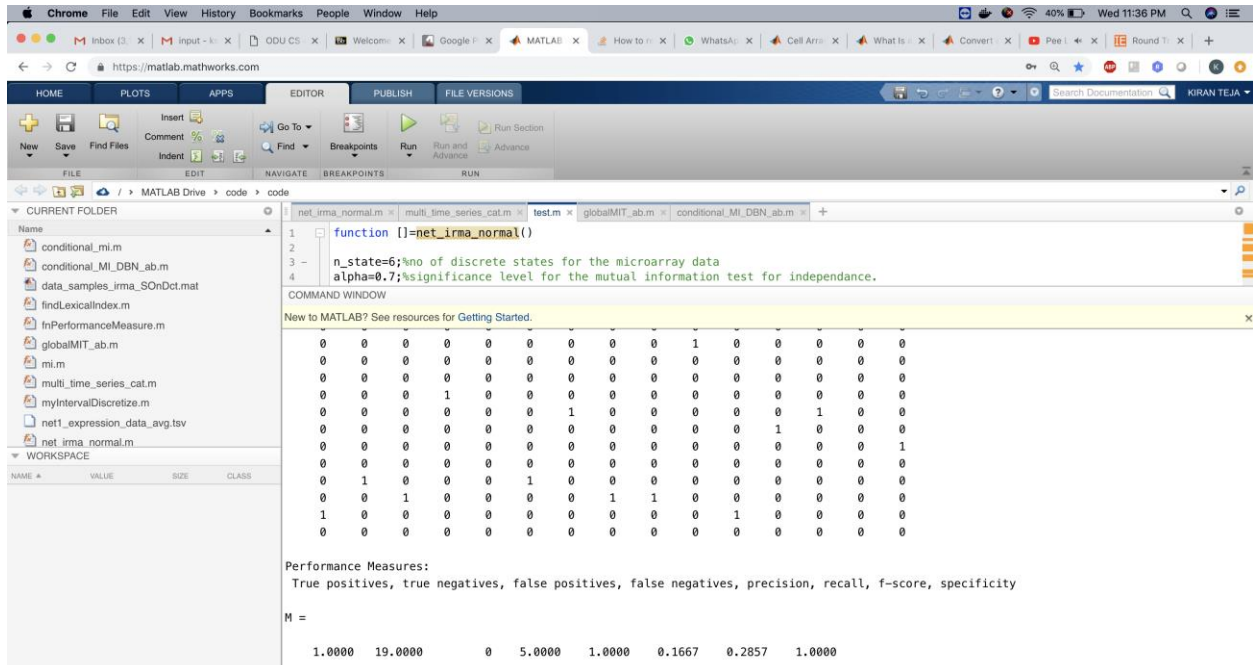
Performance Measures:
True positives, true negatives, false positives, false negatives, precision, recall, f-score, specificity

M =

1.0000	16.0000	3.0000	5.0000	0.2500	0.1667	0.2000	0.8421
--------	---------	--------	--------	--------	--------	--------	--------

>>

Best network for 10 Genes:



The screenshot shows the MATLAB Editor with the following code and output:

```
function []=net_irma_normal()
n_state=6;%no of discrete states for the microarray data
alpha=0.7;%significance level for the mutual information test for independence.
```

COMMAND WINDOW

New to MATLAB? See resources for Getting Started.

best_net =

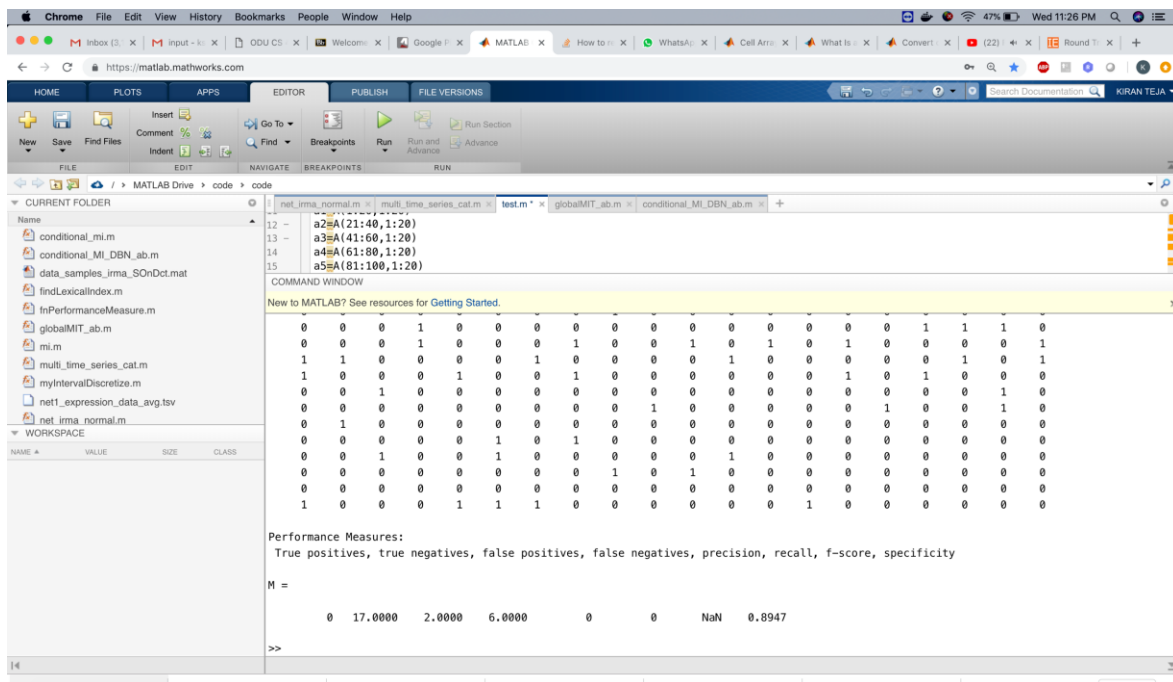
0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	1	0	0	0	0	0	0	1	0
0	0	0	0	0	0	0	0	0	0	0	0	1	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	1	0	0	0	0	1	0	0	0	0	0	0	0	0
0	0	1	0	0	0	0	1	1	0	0	0	0	0	0
1	0	0	0	0	0	0	0	0	0	1	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Performance Measures:
True positives, true negatives, false positives, false negatives, precision, recall, f-score, specificity

M =

1.0000	19.0000	0	5.0000	1.0000	0.1667	0.2857	1.0000
--------	---------	---	--------	--------	--------	--------	--------

Best network for 20 Genes:



Best network for 35 Genes:



Best network for 50 Genes:

[illegible]

Best network for 60 Genes:

[illegible]

CONCLUSION:

We studied a novel method, namely LBN, to improve the accuracy of GRN inference from gene expression data by adopting two strategies, i.e., the network decomposition and the false-positive edge deletion, which can accurately infer a directed network with high computational efficiency. Specifically, the network decomposition can effectively reduce the high computational cost of BN method for inferring large-scale GRNs, whereas CMI with kNN can delete the redundant regulations and thus reduce the false positives. By iteratively performing CMI and BN with kNN methods, LBN algorithm helps to infer the optimal GRN structure with regulation directions. We performed the experiment initially with 10 genes and then gradually scaled it to 60 genes and documented the results.

ACKNOWLEDGEMENT:

We Sincerely thank Dr. Sun Jiangwen for the sharing his knowledge with us and giving guidance, also being available all time helping us solving issues.

CITATIONS:

- <https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1005024>
- Friedman N, Linial M, Nachman I, Pe'Er D. Using Bayesian networks to analyze expression data. *Journal of Computational Biology A Journal of Computational Molecular Cell Biology*. 2000;7(3–4):601–20. pmid:11108481
- Savoie CJ, Aburatani S, Watanabe S, Eguchi Y, Muta S, Imoto S, et al. Use of gene networks from full genome microarray libraries to identify functionally relevant drug-affected genes and gene regulation cascades. *Dna Research*. 2003;10(1):19–25. pmid:12693551
- Levine M, Davidson EH. Gene regulatory networks for development. *Proc Natl Acad Sci USA. Proceedings of the National Academy of Sciences*. 2005;102(14):4936–42. pmid:15788537
- Shmulevich I, Dougherty ER, Kim S, Zhang W. Probabilistic Boolean Networks: a rule-based uncertainty model for gene regulatory networks. *Bioinformatics*. 2002;18(2):261–74. pmid:11847074